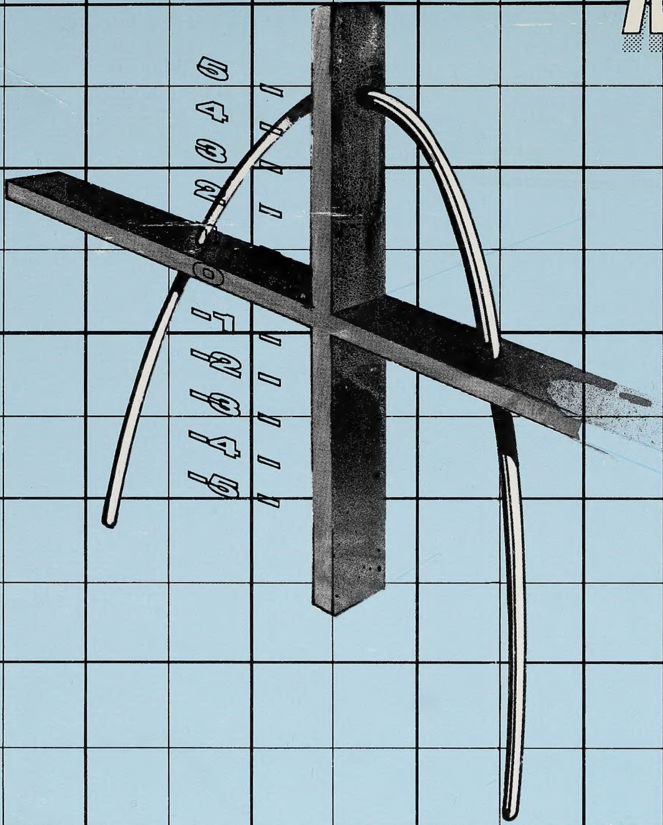
MATHEMATICS 3

STATISTICS
UNIT 6



Distance Learning

Alberta
EDUCATION

Welcome

Distance Learning

You have chosen an alternate form of learning that allows you to work at your own pace. You will be responsible for your own schedule, for disciplining yourself to study the units thoroughly, and for completing your units regularly. We wish you much success and enjoyment in your studies.

Mathematics 30 Student Module Unit 6 Statistics Alberta Correspondence School ISBN No. 0-7741-0098-2

Care has been taken to trace ownership of copyright material. Any information which will enable Alberta Education to rectify any reference or credit in subsequent printings will be gladly received by the Director, Alberta Correspondence School, Box 4000, Barrhead, Alberta, T0G 2P0.

General Information

This information explains the basic layout of each booklet.

- **What You Already Know** and **Review** are to help you look back at what you have previously studied. The questions are to jog your memory and to prepare you for the learning that is going to happen in this unit.

- As you begin each **Topic**, spend a little time looking over the components. Doing this will give you a preview of what will be covered in the topic and will set your mind in the direction of learning.

- **Exploring the Topic** includes the objectives, concept development, and activities for each objective. Use your own papers to arrive at the answers in the activities.

- **Extra Help** reviews the topic. If you had any difficulty with **Exploring the Topic**, you may find this part helpful.

- **Extensions** gives you the opportunity to take the topic one step further.

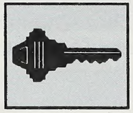
- To summarize what you have learned, and to find instructions on doing the unit assignment, turn to the **Unit Summary** at the end of the unit.

- The **Appendices** include the solutions to **Activities (Appendix A)** and any other charts, tables, etc. which may be referred to in the topics **(Appendix B, etc.)**.

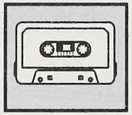
Visual Cues

Visual cues are pictures that are used to identify important areas of the material. They are found throughout the booklet.

An explanation of what they mean is written beside each visual cue.

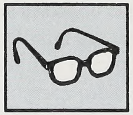


Key Idea
- flagging important ideas

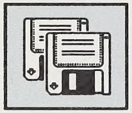


Audiotape
- learning by listening to an audiotape

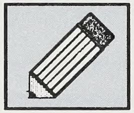
What You Already Know
- reviewing what you already know



Another View
- exploring different perspectives

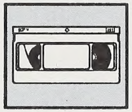


Computer Software
- learning by using computer software



Review
- studying previous concepts

Solutions
- correcting the activities

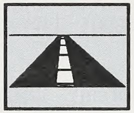


Videotape
- learning by viewing a videotape

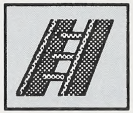
Introduction
- introducing the unit

Extra Help
- providing additional study

Print Pathway
- choosing a print alternative



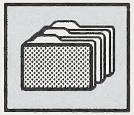
What Lies Ahead
- previewing the unit



Extensions
- going on with the topic

Calculator
- using your calculator

Exploring the Topic
- actively learning new concepts



What You Have Learned
- summarizing what you have learned

Mathematics 30

Course Overview

Mathematics 30 contains 7 units. Beside each unit is a percentage that indicates what the unit is worth in relation to the rest of the course. The units and their percentages are listed below. You will be studying the unit that is shaded.

Unit 1 Polynomial Functions	12%
Unit 2 Logarithms	10%
Unit 3 Sequences, Series, Limits	13%
Unit 4 Trigonometry	15%
Unit 5 Quadratic Relations	20%
Unit 6 **Statistics**	**18%**
Unit 7 Permutations and Combinations	12%
	100%

Unit Assessment

After completing the unit you will be given a mark based totally on a unit assignment. This assignment will be found in the Assignment Booklet.

Unit Assignment - 100%

If you are working on a CML terminal your teacher will determine what this assessment will be. It may be

Unit assignment - 50%
Supervised unit test - 50%

Introduction to Statistics

This unit covers topics dealing with statistics. Each topic contains explanations, examples, and activities to assist you in understanding statistics. If you find you are having difficulty with the explanations and the way the material is presented, there is a section called **Extra Help**. If you would like to extend your knowledge of the topic, there is a section called **Extensions**.

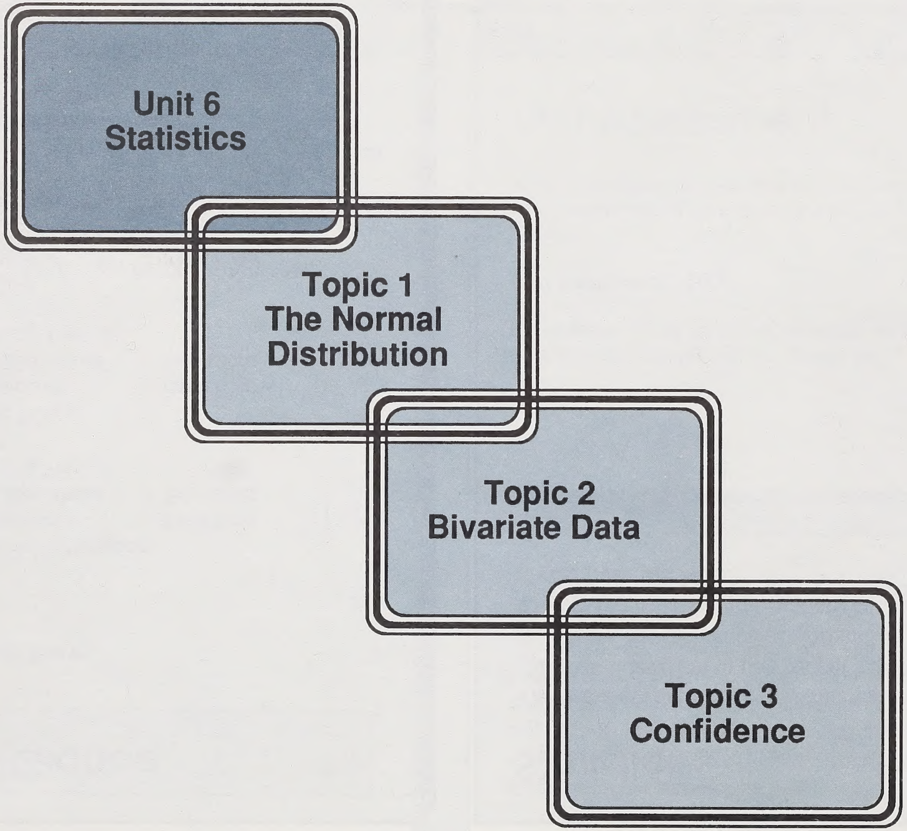
You can evaluate your understanding of each topic by working through the activities. Answers are found in the Solutions in **Appendix A.** In several cases there is more than one way to do the question.

Unit 6 Statistics

Contents at a Glance

Statistics

The word **statistics** is everywhere these days. From the football game on television to newspaper reports on industries like insurance, the influence of statistics is ever present. This unit attempts to broaden your understanding and separate some myths from the facts.



Unit 6
Statistics

Topic 1
The Normal
Distribution

Topic 2
Bivariate Data

Topic 3
Confidence

What You Already Know

Refresh your memory!

Do you remember how to do the following?

1. Identify the terms data, sample, population (universe), measure of central tendency, mean, median, mode, discrete data, continuous data, range, standard deviation, frequency, class, class boundaries or class limits, class mark, class width, and frequency distribution table.

2. Determine the range of a set of grouped or ungrouped data.

3. Determine the mean, median, and mode of a set of grouped or ungrouped data.

4. Present an argument for choosing a particular measure of central tendency for a set of grouped or ungrouped data.

5. Interpret statistical data given in graphical form.

6. Organize a set of data into a frequency distribution table which includes class boundaries, class mark, and frequency for each class.

7. Given a set of data, determine the mean.

8. Given a set of data, determine the median.

9. Given two points on a line, determine the slope of the line.

10. Given two points on a line, determine the equation of the line.

11. Given an ordered pair(s), plot the ordered pair(s) on a Cartesian plane.

12. Given a line, estimate the ordered pairs of two points on the line.

13. Given a ratio, determine the equivalent percentage.

Now that you have looked at material that you studied previously, go to the **Review** to confirm your understanding of this material.

Review

Try the following review questions.

1. Match by placing the appropriate letter in the blank provided.

a. data
b. sample
c. population
d. measures of central tendency
e. mean
f. mode
g. median
h. discrete data
i. continuous data
j. measure of dispersion
k. range
l. frequency
m. class
n. class limit
o. class mark
p. class width
q. frequency
r. frequency distribution table

_____ the set of people or things on which information is required

_____ the average of the data

_____ the difference between the highest and lowest values in the data

_____ measures of how the data is spread

_____ facts or information

_____ the distance between the lower class boundaries of two adjacent classes

_____ the value that appears most often in the data

_____ data that is countable

_____ the number of values within each class

_____ the upper and lower values between which a piece of data must fit

_____ the midpoint of a class

2. Determine the range for the following two sets of data.

a.

346	325	295	402	386	354	347	336
389	352	409	372	361	356	409	398
409	355	376	382	345	289	366	349

b.

Class	Class Limits	Frequency
1	6 - 11	5
2	12 - 17	11
3	18 - 23	23
4	24 - 29	45
5	30 - 35	27
6	36 - 41	13
7	42 - 47	2

3. Find the measures of central tendency for the following sets of data.

a.

346	325	295	402	386	354	347	336
389	352	409	372	361	356	409	398
409	355	376	382	345	289	366	349

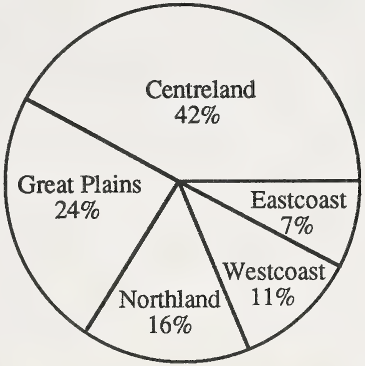
b.

Class	Class Limits	Frequency
1	6 - 11	5
2	12 - 17	11
3	18 - 23	23
4	24 - 29	45
5	30 - 35	27
6	36 - 41	13
7	42 - 47	2

4. In question 3a, which measure of central tendency best describes the centre of the data? Explain your answer.

5. Use the following graph to answer the questions.

The Number of People
Employed in Eustania
Total Employment: 15 000 000



a. What percentage of the people work in Westcoast?

b. How many people work in Centreland?

c. Which area has more employees: Centreland or the Great Plains, Westcoast, and Eastcoast combined?

d. How many more people work in the Great Plains than in Northland?

6. Construct a frequency distribution table using the following information. Make sure the table has classes, class limits, class mark, and frequency.

346	325	295	402	386	354	347	336
389	352	409	372	361	356	409	398
409	355	376	382	345	289	366	349
327	387	356	349	368	407	287	354
348	359	348	362	337	372	392	315
367	362	373	359	352	332	354	361

7. Find the mean for the following groups of data.

a. 18, 34, 14

b. 6, 4, 7, 10, 9, 9, 11, 16

c. 141, 107, 118, 114, 125

8. Find the median for the following groups of data.

a. 18, 34, 14 b. 6, 4, 7, 10, 9, 9, 11, 16

c. 141, 107, 118, 114, 125

9. Determine the slope of a line that passes through each of the following pairs of points.

a. (4, 7) and (9, 5) b. (108, 22) and (112, 102)

10. Determine the equation of the line that passes through each of the following pairs of points.

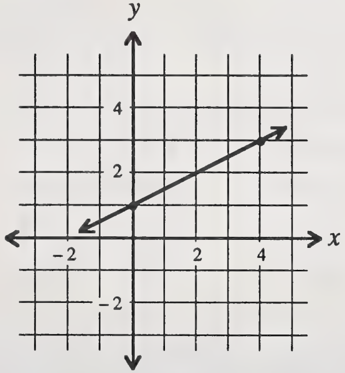
a. (4, 7) and (9, 5) b. (108, 22) and (112, 102)

11. Plot the following ordered pairs on a Cartesian plane.

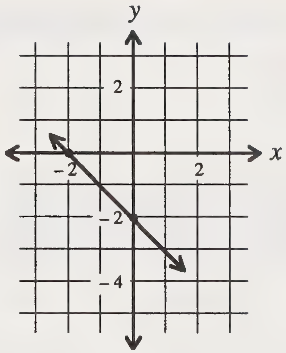
a. $A\,(3, 4)$ b. $B\,(-2, 5)$ c. $C\,(-5, 0)$

12. Find the slope of the following lines.

a.



b.



13. Convert the following ratios to percent.

a. $\frac{12}{25}$

b. $\frac{17}{20}$

c. $\frac{9}{60}$

✓ Now go to the **Review** solutions in **Appendix A.**

Topic 1 The Normal Distribution

Introduction

It is a long and tedious task working through all the calculations which find the measures and aspects of a set of data.

If a set of data has the same characteristics as a set of data that you have worked with before, you can use the results from the previous set to lessen your work.

This is exactly what has been done for sets of data that are normally distributed.

In this topic, you will see how all normal distributions are related and how these similarities can be used to save you the task of doing long and boring calculations.

This will lead you to the use of *z*-scores which will permit you to use probability to make predictions about populations.

What Lies Ahead

Throughout this topic you will learn to

1. calculate and interpret the mean and the standard deviation of a set of data
2. identify the normal distribution
3. use *z*-scores to solve situations that are normally distributed
4. use *z*-scores to calculate the probability of an event happening

Now that you know what to expect, turn the page to begin your study of the normal distribution.

Exploring Topic 1

Activity 1

Calculate and interpret the mean and the standard deviation of a set of data.

Standard deviation is one of the most frequently used measures of dispersion. This measure describes the distribution of data about the mean.

Now you will use a particular case to learn how to calculate standard deviation.

A small group of investors is planning to build a ski resort. The group has found only three locations that meet its criteria. The decision on which area to purchase is going to be based on the amount of snowfall on the slopes. The amount of snowfall, in centimetres, on each of the slopes over the last fifteen years follows. Based on this data, which site will the group select?

Sunridge Site

305, 237, 348, 373, 279,
262, 331, 339, 271, 357,
253, 219, 391, 336, 274

Glacier Site

292, 255, 307, 318, 353,
281, 342, 329, 268, 289,
321, 312, 298, 319, 291

Alpine Basin Site

321, 267, 232, 304, 289,
342, 380, 403, 207, 294,
316, 237, 373, 251, 359

The standard deviation will be calculated using a different instrument for each site. For the Sunridge site, the calculations will be done using pencil and paper. This method most clearly shows how standard deviation is calculated. An alternate method of using a chart will also be shown beside the calculations.

Step 1: Calculate the mean for the data.

The mean is found by dividing the sum of the data values by the number of pieces of data.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{305 + 237 + 348 + \ . \ . \ . \ + 274}{15}$$

$$\bar{x} = \frac{4575}{15}$$

$$\bar{x} = 305$$

Value (x_i)	Mean $(\bar{x})$
305	305
237	305
348	305
373	305
279	305
262	305
331	305
339	305
271	305
357	305
253	305
219	305
391	305
336	305
274	305

Step 2: Subtract the mean $(\bar{x})$ from each data value. These numbers are called the **deviations** from the mean.

$$\begin{aligned} x_1 - \bar{x} &= 305 - 305 \\ &= 0 \end{aligned}$$

$$\begin{aligned} x_2 - \bar{x} &= 237 - 305 \\ &= -68 \end{aligned}$$

$$\begin{aligned} x_3 - \bar{x} &= 348 - 305 \\ &= 43 \end{aligned}$$

$$\begin{aligned} x_4 - \bar{x} &= 373 - 305 \\ &= 68 \end{aligned}$$

$$\begin{aligned} x_5 - \bar{x} &= 279 - 305 \\ &= -26 \end{aligned}$$

$$\begin{aligned} x_6 - \bar{x} &= 262 - 305 \\ &= -43 \end{aligned}$$

$$\begin{aligned} x_7 - \bar{x} &= 331 - 305 \\ &= 26 \end{aligned}$$

$$\begin{aligned} x_8 - \bar{x} &= 339 - 305 \\ &= 34 \end{aligned}$$

$$\begin{aligned} x_9 - \bar{x} &= 271 - 305 \\ &= -34 \end{aligned}$$

$$\begin{aligned} x_{10} - \bar{x} &= 357 - 305 \\ &= 52 \end{aligned}$$

$$\begin{aligned} x_{11} - \bar{x} &= 253 - 305 \\ &= -52 \end{aligned}$$

$$\begin{aligned} x_{12} - \bar{x} &= 219 - 305 \\ &= -86 \end{aligned}$$

$$\begin{aligned} x_{13} - \bar{x} &= 391 - 305 \\ &= 86 \end{aligned}$$

$$\begin{aligned} x_{14} - \bar{x} &= 336 - 305 \\ &= 31 \end{aligned}$$

$$\begin{aligned} x_{15} - \bar{x} &= 274 - 305 \\ &= -31 \end{aligned}$$

Value (x_i)	Mean $(\bar{x})$	Deviation $(x_i - \bar{x})$
305	305	0
237	305	−68
348	305	43
373	305	68
279	305	−26
262	305	−43
331	305	26
339	305	34
271	305	−34
357	305	52
253	305	−52
219	305	−86
391	305	86
336	305	31
274	305	−31

Value (x_i)	Mean $(\bar{x})$	Deviation $(x_i - \bar{x})$	Deviation^2 $(x_i - \bar{x})^2$
305	305	0	0
237	305	−68	4624
348	305	43	1849
373	305	68	4624
279	305	−26	676
262	305	−43	1849
331	305	26	676
339	305	34	1156
271	305	−34	1156
357	305	52	2704
253	305	−52	2704
219	305	−86	7396
391	305	86	7396
336	305	31	961
274	305	−31	961

Step 3: Square the deviations that you calculated in the last step.

Squaring removes the negative signs from the deviations and makes the measures more sensitive to small changes.

$$\begin{aligned}(x_1 - \bar{x})^2 &= 0^2\\ &= 0\end{aligned} \qquad \begin{aligned}(x_2 - \bar{x})^2 &= (-68)^2\\ &= 4624\end{aligned}$$

$$\begin{aligned}(x_3 - \bar{x})^2 &= (43)^2\\ &= 1849\end{aligned} \qquad \begin{aligned}(x_4 - \bar{x})^2 &= (68)^2\\ &= 4624\end{aligned}$$

$$\begin{aligned}(x_5 - \bar{x})^2 &= (-26)^2\\ &= 676\end{aligned} \qquad \begin{aligned}(x_6 - \bar{x})^2 &= (-43)^2\\ &= 1849\end{aligned}$$

$$\begin{aligned}(x_7 - \bar{x})^2 &= (26)^2\\ &= 676\end{aligned} \qquad \begin{aligned}(x_8 - \bar{x})^2 &= (34)^2\\ &= 1156\end{aligned}$$

$$\begin{aligned}(x_9 - \bar{x})^2 &= (-34)^2\\ &= 1156\end{aligned} \qquad \begin{aligned}(x_{10} - \bar{x})^2 &= (52)^2\\ &= 2704\end{aligned}$$

$$(x_{11}-\bar{x})^2=(-52)^2$$
$$=2704$$

$$(x_{12}-\bar{x})^2=(-86)^2$$
$$=7396$$

$$(x_{13}-\bar{x})^2=(86)^2$$
$$=7396$$

$$(x_{14}-\bar{x})^2=(31)^2$$
$$=961$$

$$(x_{15}-\bar{x})^2=(-31)^2$$
$$=961$$

Step 4: Calculate the mean of the square deviations which were found in Step 3.

Value (x_i)	Mean $(\bar{x})$	Deviation $(x_i-\bar{x})$	Deviation^2 $(x_i-\bar{x})^2$
305	305	0	0
237	305	−68	4624
348	305	43	1849
373	305	68	4624
279	305	−26	676
262	305	−43	1849
331	305	26	676
339	305	34	1156
271	305	−34	1156
357	305	52	2704
253	305	−52	2704
219	305	−86	7396
391	305	86	7396
336	305	31	961
274	305	−31	961

The sum of values in the fourth column is divided by the number of values. This quotient is referred to as the **variance**. Variance is another name for mean square deviation.

$$\text{mean square deviation}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$
$$=\frac{4624+1849+4624+\ .\ .\ .\ +961}{15}$$
$$=\frac{38\ 732}{15}$$
$$\doteq 2582.1$$

Step 5: Take the square root of the mean square deviation. This is the **standard deviation** of the data. The symbol that represents standard deviation is σ.

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$
$$\sigma=\sqrt{2582.1}$$
$$\sigma\doteq 50.8$$

The standard deviation of the snowfall on the Sunridge site is 50.8 cm.

Using the chart instead of writing out all of the calculations saves you a lot of time and all of the information is still written down. Another method that saves you a considerable amount of time is the calculator.

Many calculators now contain statistical functions, but you must be careful when you are using the calculator since you will not be permitted to see the information that you have entered into the calculator. The calculator is designed to give you only the answers to the different types of functions.

If your calculator does not have a statistical function, move on to the next method.

The procedure for only one type of calculator is shown in this activity. Check the owner's manual for your calculator to see the correct method for using your calculator.

Now a calculator will be used to calculate the standard deviation for the Glacier site.

Step 1: Turn on the calculator's statistical mode.

Press the $\boxed{\text{MODE}}$ key followed by the $\boxed{\cdot}$ key. This will turn on the statistical mode of the calculator. An SD will appear in the upper right corner of the display to inform you that the calculator is in its statistical mode.

Step 2: Press $\boxed{\text{INV}}$ followed by the $\boxed{\text{SAC}}$ key.

This informs the calculator that you are going to enter the values of the data.

Step 3: Enter the values, followed by the $\boxed{x}$ key after each piece of data.

292 $\boxed{x}$ 255 $\boxed{x}$ 307 $\boxed{x}$ 318 $\boxed{x}$ 353 $\boxed{x}$

281 $\boxed{x}$ 342 $\boxed{x}$ 329 $\boxed{x}$ 268 $\boxed{x}$ 289 $\boxed{x}$

321 $\boxed{x}$ 312 $\boxed{x}$ 298 $\boxed{x}$ 319 $\boxed{x}$ 291 $\boxed{x}$

Step 4: The standard deviation is then found by pressing the $\boxed{\text{INV}}$ key followed by the $\boxed{\sigma n}$ key.

Pressing the $\boxed{\text{INV}}$ and $\boxed{\sigma n}$ keys will give the number 25.868 900 25 which is then rounded to 25.9.

The words written inside a box will be the words or symbols that appear on or above the keys.

The following keys can also be used to perform different calculations.

mean	INV	$\bar{x}$	305
number of values	INV	n	15
sum of values	INV	Σx	4575

Another procedure that can be used is a spreadsheet on a microcomputer. This method gives you the speed of a calculator and a detailed chart like the chart used in the first method.

For the Alpine Basin site, the *AppleWorks* ™ [1] spreadsheet will be used since it is recognized as a recommended resource by Alberta Education.

If you have a different spreadsheet, check your manual for any differences in the commands that you will have to make to use the spreadsheet.

If you do not have a spreadsheet or microcomputer, you may skip over this method.

Step 1: Place the values of the data in the first column of the spreadsheet. Place the label for the values (x) in the top cell.

```
======A======B======C======D======E======
  1| x
  2|     321
  3|     267
  4|     232
  5|     304
  6|     289
  7|     342
  8|     380
  9|     403
 10|     207
 11|     294
 12|     316
 13|     237
 14|     373
 15|     251
 16|     359
 17|
 18|
------------------------------------------------------------
```

You should be able to use one of the first two methods since you will not be permitted to use a microcomputer while writing your exam.

This section assumes that you already know how to use the *AppleWorks* ™ [1] spreadsheet. If you do not, please read your *AppleWorks* ™ [1] instruction manual before doing this method.

[1] *AppleWorks* ™ is a trademark of Apple Computers, Inc.

Step 2: In cell B1, write down the label for the second column, $\bar{x}$. In cell B2, write down the command @AVG(A2 . . . A16). Copy this cell into the cells below through to the end of the sixteenth row.

This is the command that tells the microcomputer to take the mean of all of the numbers in the first column.

```
======A======B======C======D======E======
  1| x         mean
  2|    321      305
  3|    267      305
  4|    232      305
  5|    304      305
  6|    289      305
  7|    342      305
  8|    380      305
  9|    403      305
 10|    207      305
 11|    294      305
 12|    316      305
 13|    237      305
 14|    373      305
 15|    251      305
 16|    359      305
 17|
 18|
-------------------------------------------------------------------
```

Step 3: In the first cell of the third column, write down the label **x – mean.** In cell C2, write down the formula +A2 – B2. Make a relative copy of this cell to the rest of the cells in this column.

The relative copy of this formula will change the number in the cell location for each row. For example, in the fifth row, the formula will become +A5 – B5.

If you have copied the formula properly, there should be fifteen different numbers in this column.

```
======A======B======C======D======E======
  1| x          mean   x – mean
  2|    321      305       16
  3|    267      305     – 38
  4|    232      305     – 73
  5|    304      305      – 1
  6|    289      305     – 16
  7|    342      305       37
  8|    380      305       75
  9|    403      305       98
 10|    207      305     – 98
 11|    294      305     – 11
 12|    316      305       11
 13|    237      305     – 68
 14|    373      305       68
 15|    251      305     – 54
 16|    359      305       54
 17|
 18|
-------------------------------------------------------------------
```

Step 4: In the first cell of the fourth column, write in the label **sq devi**. In cell D2, write the formula +(C2)^2. Make a relative copy of this formula into the rest of the cells of this column.

This formula takes the squares of the deviations from the mean which are in the third column.

```
======A======B======C======D======E======

 1| x        mean   x – mean  sq devi
 2|    321    305       16      256
 3|    267    305     – 38     1444
 4|    232    305     – 73     5329
 5|    304    305      – 1        1
 6|    289    305     – 16      256
 7|    342    305       37     1369
 8|    380    305       75     5625
 9|    403    305       98     9604
10|    207    305     – 98     9604
11|    294    305     – 11      121
12|    316    305       11      121
13|    237    305     – 68     4624
14|    373    305       68     4624
15|    251    305     – 54     2916
16|    359    305       54     2916
17|
18|
--------------------------------------------------------------
```

Step 5: In cell D17, write the formula @SUM(D2 . . . D16). In cell E17, write the formula +(D17/15)^.5. Write the label **standard deviation** in cell E16.

The first command will calculate the sum of the squares of the deviations, while the second command will take the square root of the mean of the square deviations. This second command will give you the standard deviation for the data.

```
======A======B======C======D======E======

 1| x        mean   x – mean  sq devi
 2|    321    305       16      256
 3|    267    305     – 38     1444
 4|    232    305     – 73     5329
 5|    304    305      – 1        1
 6|    289    305     – 16      256
 7|    342    305       37     1369
 8|    380    305       75     5625
 9|    403    305       98     9604
10|    207    305     – 98     9604
11|    294    305     – 11      121
12|    316    305       11      121
13|    237    305     – 68     4624
14|    373    305       68     4624
15|    251    305     – 54     2916
16|    359    305       54     2916 standard deviation
17|                           48810 57.04384
18|
--------------------------------------------------------------
```

The standard deviation for this group of data will be rounded to 57.0.

Summarize the standard deviations that were calculated for each of the sites.

Site	Mean (cm)	Standard Deviation (cm)
Sunridge	305	50.8
Glacier	305	25.9
Alpine Basin	305	57.0

How is the standard deviation used to select the better site for the resort?

As noted at the start of this activity, the standard deviation tells you how the data is distributed around the mean. The spread for the Glacier site is smaller than either of the spreads for the Sunridge and Alpine Basin sites. This means that the snowfall for the Glacier site is more consistent than for the other two sites.

The typical snowfall for the three sites will be between the following measures.

Sunridge (305 – 50.8) cm and (305 + 50.8) cm
or 254.2 cm and 355.8 cm

Glacier (305 – 25.9) cm and (305 + 25.9) cm
or 279.1 cm and 330.9 cm

Alpine Basin (305 – 57.0) cm and (305 + 57.0) cm
or 248 cm and 362 cm

In those years when there will be less snowfall, the Glacier site will receive more snow. Therefore, the group would select the Glacier site.

In this last example, all of the values in the data were given. There will be times when you will be given a table of values instead of the actual values.

Look at how you will find the standard deviation in those cases.

The following tables show the number of skis of each length requested for rental by customers. The standard deviation will be determined for the pattern of renting the different lengths of skis.

Length: 170 cm

Class Number	Number of Requests	Number of Days
1	50 - 59	17
2	60 - 69	27
3	70 - 79	49
4	80 - 89	72
5	90 - 99	44
6	100 - 109	29
7	110 - 119	12

Length: 180 cm

Class Number	Number of Requests	Number of Days
1	50 - 59	22
2	60 - 69	34
3	70 - 79	41
4	80 - 89	54
5	90 - 99	43
6	100 - 109	38
7	110 - 119	18

Length: 190 cm

Class Number	Number of Requests	Number of Days
1	50 - 59	23
2	60 - 69	21
3	70 - 79	45
4	80 - 89	55
5	90 - 99	51
6	100 - 109	28
7	110 - 119	27

There is one significant difference between these tables and the groups of data that were used before to find the standard deviation. These tables do not tell you the actual values.

A way must be developed that will permit the calculation of the standard deviation without using the actual values of the data.

What you do know is the range in which the values must fall. For example, on 17 days there were between 50 and 59 requests for 170 cm skis. You will assume that the 17 days are equally distributed between 50 and 59 requests. Therefore, for the calculations, it will be assumed that 54.5 (the average of 50 and 59) 170 cm skis were requested on 17 different days. This same assumption will be made for all of the classes.

Step 1: Find the class midpoint for each class. Class midpoint is another name for class mark.

Class 1: $\frac{50+59}{2}=\frac{109}{2}$ $=54.5$

Class 2: $\frac{60+69}{2}=\frac{129}{2}$ $=64.5$

Class 3: $\frac{70+79}{2}=\frac{149}{2}$ $=74.5$

Class 4: $\frac{80+89}{2}=\frac{169}{2}$ $=84.5$

Class 5: $\frac{90+99}{2}=\frac{189}{2}$ $=94.5$

Class 6: $\frac{100+109}{2}=\frac{209}{2}$ $=104.5$

Class 7: $\frac{110+119}{2}=\frac{229}{2}$ $=114.5$

Length: 170 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)
1	50 - 59	17	54.5
2	60 - 69	27	64.5
3	70 - 79	49	74.5
4	80 - 89	72	84.5
5	90 - 99	44	94.5
6	100 - 109	29	104.5
7	110 - 119	12	114.5

Step 2: Multiply the frequency by the class midpoint. Find the sum of these products and the sum of the frequencies. Divide the sum of the frequencies into the sum of the products. This will be the mean for the data.

Length: 170 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$
1	50 - 59	17	54.5	926.5
2	60 - 69	27	64.5	1 741.5
3	70 - 79	49	74.5	3 650.5
4	80 - 89	72	84.5	6 084
5	90 - 99	44	94.5	4 158
6	100 - 109	29	104.5	3 030.5
7	110 - 119	12	114.5	1 374
		250		20 965

$$\text{mean} = \frac{\sum(f \cdot x)}{n}$$

$$\text{mean} = \frac{20\,965}{250}$$

$$\text{mean} = 83.86$$

$$\text{mean} \doteq 84$$

Step 3: For each class subtract the mean from each class midpoint.

Length: 170 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x - \mu$
1	50 - 59	17	54.5	926.5	– 29.5
2	60 - 69	27	64.5	1 741.5	– 19.5
3	70 - 79	49	74.5	3 650.5	– 9.5
4	80 - 89	72	84.5	6 084	0.5
5	90 - 99	44	94.5	4 158	10.5
6	100 - 109	29	104.5	3 030.5	20.5
7	110 - 119	12	114.5	1 374	30.5
		250		20 965	

The symbol μ represents the mean.

Step 4: Find the square of the difference between class midpoints and the mean.

Length: 170 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$
1	50 - 59	17	54.5	926.5	– 29.5	870.25
2	60 - 69	27	64.5	1 741.5	– 19.5	380.25
3	70 - 79	49	74.5	3 650.5	– 9.5	90.25
4	80 - 89	72	84.5	6 084	0.5	0.25
5	90 - 99	44	94.5	4 158	10.5	110.25
6	100 - 109	29	104.5	3 030.5	20.5	420.25
7	110 - 119	12	114.5	1 374	30.5	930.25
		250		20 965		

Step 5: Find the product of the frequencies and the squares of the mean deviations. Find the sum of these products.

Length: 170 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	50 - 59	17	54.5	926.5	– 29.5	870.25	14 794.25
2	60 - 69	27	64.5	1 741.5	– 19.5	380.25	10 266.75
3	70 - 79	49	74.5	3 650.5	– 9.5	90.25	4 422.25
4	80 - 89	72	84.5	6 084	0.5	0.25	18
5	90 - 99	44	94.5	4 158	10.5	110.25	4 851
6	100 - 109	29	104.5	3 030.5	20.5	420.25	12 187.25
7	110 - 119	12	114.5	1 374	30.5	930.25	11 163
		250		20 965			57 702.5

Step 6: Divide the sum of the products of the frequencies and the squares of the mean deviations by the total frequency. The square root of the quotient is the standard deviation.

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{57\ 702.5}{250}}$$

$$\sigma = \sqrt{230.81}$$

$$\sigma \doteq 15.2$$

The standard deviations for the other two ski lengths will be calculated using a chart. They can also be calculated using a calculator or spreadsheet on a microcomputer.

If you would like to see how to use your calculator to do these calculations, turn to **Extensions** at the end of this topic.

Length: 180 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	50 - 59	22	54.5	1 199	– 29.9	894.01	19 668.22
2	60 - 69	34	64.5	2 193	– 19.9	396.01	13 464.34
3	70 - 79	41	74.5	3 054.5	– 9.9	98.01	4 018.41
4	80 - 89	54	84.5	4 563	0.1	0.01	0.54
5	90 - 99	43	94.5	4 063.5	10.1	102.01	4 386.43
6	100 - 109	38	104.5	3 971	20.1	404.01	15 352.38
7	110 - 119	18	114.5	2 061	30.1	906.01	16 308.18
		250		21 105			73 198.5

$$\text{mean} = \frac{\sum (f \cdot x)}{n}$$

$$\text{mean} = \frac{21\,105}{250}$$

$$\text{mean} = 84.42$$

$$\text{mean} \doteq 84.4$$

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{73\,198.5}{250}}$$

$$\sigma = \sqrt{292.794}$$

$$\sigma \doteq 17.1$$

Length: 190 cm

Class Number	Class Limits	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	50 - 59	23	54.5	1 253.5	– 31.3	979.69	22 532.87
2	60 - 69	21	64.5	1 354.5	– 21.3	453.69	9 527.49
3	70 - 79	45	74.5	3 352.5	– 11.3	127.69	5 746.05
4	80 - 89	55	84.5	4 647.5	– 1.3	1.69	92.95
5	90 - 99	51	94.5	4 819.5	8.7	75.69	3 860.19
6	100 - 109	28	104.5	2 926.0	18.7	349.69	9 791.32
7	110 - 119	27	114.5	3 091.5	28.7	823.69	22 239.63
		250		21 445.0			73 790.50

$$\text{mean} = \frac{\sum (f \cdot x)}{n} \qquad \sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\text{mean} = \frac{21\,445.0}{250} \qquad \sigma = \sqrt{\frac{73\,790.50}{250}}$$

$$\text{mean} \doteq 85.8 \qquad \sigma = \sqrt{295.162}$$

$$\sigma \doteq 17.2$$

The smallest standard deviation is usually the most useful since it means the data is less dispersed about the mean.

Answer any five of the following questions.

1. Explain what is meant by dispersion.

2. A golfer got the following scores in the last twelve rounds.

102	97	105	105	96	111
106	100	99	103	102	98

Calculate the following quantities for the golfer.

a. μ b. $\sum(x-\mu)$

c. $\sum(x-\mu)^2$ d. σ

3. The masses (g) of potato chips packaged by two different companies are given below. Which company is more consistent in its packaging process?

New Duetch				Hosts			
198	205	201	200	199	201	201	197
207	192	194	199	201	198	196	203
206	210	198	193	198	206	203	196
195	192	207	203	206	197	202	196

4. Describe the set of data that has a standard deviation of zero.

5. The following chart represents the number of motorcycles sold on various days of the month of July for the Rev Up and Go dealership.

Class	Number Sold	Frequency
1	1 - 3	2
2	4 - 6	3
3	7 - 9	5
4	10 - 12	10
5	13 - 15	6
6	16 - 18	3
7	19 - 21	2

Calculate the following quantities.

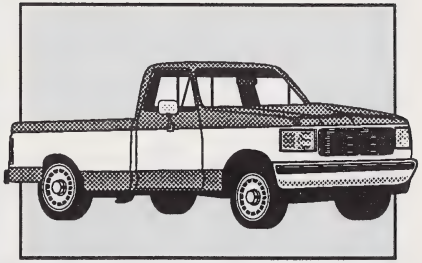
a. μ b. $\sum(x-\mu)$

c. $\sum(x-\mu)^2$ d. σ

6. The following two charts show the amount premiums have increased for two automobile insurance companies.

The Big Stone		
Class	Percent Increase	Frequency
1	0 - 0.99	3
2	1 - 1.99	5
3	2 - 2.99	9
4	3 - 3.99	7
5	4 - 4.99	3
6	5 - 5.99	3

Royal Eastern		
Class	Percent Increase	Frequency
1	0 - 0.99	3
2	1 - 1.99	6
3	2 - 2.99	7
4	3 - 3.99	7
5	4 - 4.99	4
6	5 - 5.99	3



a. Calculate the standard deviation for each company.

b. Which company has the most consistent increases in premiums?

c. Which company has the largest sum for the $f(x-\mu)^2$ column?

✓ For solutions to **Activity 1**, turn to **Appendix A, Topic 1**.

Activity 2

Identify the normal distribution.

As you have seen in the first activity, calculating the mean and standard deviation is very time-consuming. For the sake of the explanations the number of pieces of data involved in the preceding calculations was kept small, from eight to sixteen pieces. What happens when you have to deal with a set that has thousands or millions of pieces of data?

A new method has to be developed to deal with these situations.

Take a look at two different situations.

The data collected in the first case represents the number of skis sold by a small ski shop.

Case 1

Ski Length (cm)	Number Sold
140	1
150	8
160	28
170	56
180	70
190	56
200	28
210	8
220	1

The data collected in the second case represents the students' heights in a large high school.

Case 2

Height (cm)	Number of Students
150 - 162	5
163 - 175	40
176 - 188	140
189 - 201	280
202 - 214	350
215 - 227	280
228 - 240	140
241 - 253	40
254 - 266	5

Now find the standard deviation for both of these cases.

Case 1

Skis Sold

Ski Length (cm) (x)	Number Sold (f)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
140	1	140	– 40	1600	1 600
150	8	1 200	– 30	900	7 200
160	28	4 480	– 20	400	11 200
170	56	9 520	– 10	100	5 600
180	70	12 600	0	0	0
190	56	10 640	10	100	5 600
200	28	5 600	20	400	11 200
210	8	1 680	30	900	7 200
220	1	220	40	1600	1 600
	256	46 080			51 200

$$\mu = \frac{\sum (f \cdot x)}{n} \qquad \sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\mu = \frac{46\,080}{256} \qquad \sigma = \sqrt{\frac{51\,200}{256}}$$

$$\mu = 180 \qquad \sigma \doteq 14.1$$

Case 2

Students' Heights

Height (cm)	Number (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
150 - 162	5	156	780	– 52	2704	13 520
163 - 175	40	169	6 760	– 39	1521	60 840
176 - 188	140	182	25 480	– 26	676	94 640
189 - 201	280	195	54 600	– 13	169	47 320
202 - 214	350	208	72 800	0	0	0
215 - 227	280	221	61 880	13	169	47 320
228 - 240	140	234	32 760	26	676	94 640
241 - 253	40	247	9 880	39	1521	60 840
254 - 266	5	260	1 300	52	2704	13 520
	1280		266 240			432 640

$$\mu = \frac{\sum (f \cdot x)}{n}$$

$$\mu = \frac{266\ 240}{1280}$$

$$\mu = 208$$

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{432\ 640}{1280}}$$

$$\sigma \doteq 18.4$$

Do you see a connection between the two standard deviations? No, there is no connection between student height and the length of skis they use.

In each case, use the standard deviation as a measuring unit.

What percentage of the frequency falls within one standard deviation of the mean?

Case 1

The mean length of the skis sold is 180 cm. Therefore, the values must fall between 165.9 cm (180 cm – 14.1 cm) and 194.1 cm (180 cm + 14.1 cm).

How many skis fall into this range?

Ski Length (cm)	Number Sold
170	56
180	70
190	56
	Total 182

182 skis fall into this range.

What percentage is this of the total number of skis sold? The shop sold a total of 256 skis.

$$\begin{aligned}\text{Percentage} &= \frac{182}{256} \times 100\% \\ &\doteq 71\%\end{aligned}$$

Case 2

The mean height of the students is 208 cm. Therefore, the range of the values is 189.6 cm (208 cm – 18.4 cm) and 226.4 cm (208 cm + 18.4 cm).

How many students fall into this range?

Use all of the values within the two outside intervals since there is only a very small fraction not included in the intervals.

Height (cm)	Number of Students
189 - 201	280
202 - 214	350
215 - 227	280
	Total 910

910 students fall into this range.

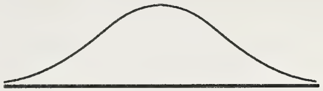
What percentage is this of the total number of students? The school has a total student population of 1280 students.

$$\begin{aligned}\text{Percentage} &= \frac{910}{1280} \times 100\% \\ &\doteq 71\%\end{aligned}$$

Now you have a connection between the two sets of data. Both sets have the same percentage of data within one standard deviation of the mean.

If you checked to see what percentage of the values was within two standard deviations and three standard deviations of the mean, they would still be the same. **Note that the method shown for Cases 1 and 2 is only approximate.**

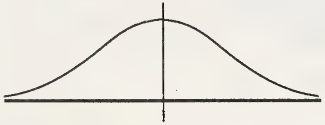
If you made a frequency polygon (ski length versus frequency and height versus frequency) of both of these sets of data, they would have the same shape, that is, the bell shape or the normal distribution curve. This shape is shown here.



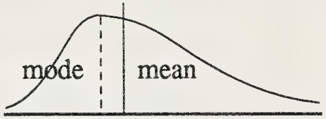
The shape of this graph is also typical of many other sets of data. All sets of data with graphs of this shape are said to be **normally distributed** or they are called **normal distributions**.

Look at the characteristics that are present in all normal distributions.

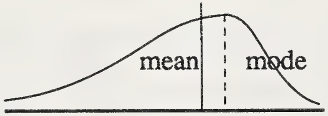
- The mean, mode, and median are all at the middle of the graph.



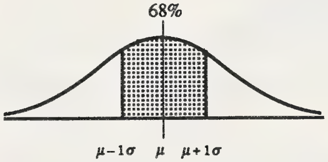
A graph like the following is said to be skewed to the left. For this case, the mode is to the left of the mean.

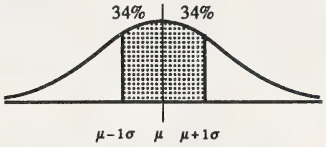


The following graph is said to be skewed to the right. Here the mode is to the right of the mean.



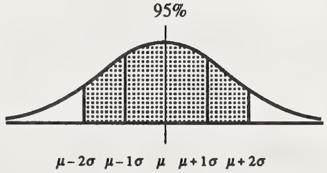
- The percentage of the total area (the percentage of the total value of data) underneath the graph within one standard deviation of the mean is 68%.
  There is 34% on each side of the mean.

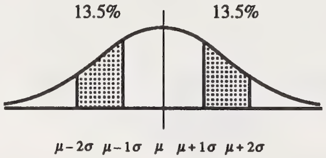




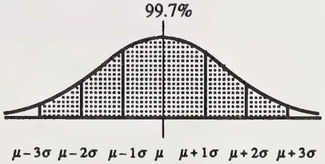
The sets of data dealing with ski length and student height were 71%, since they were small sets. As sets get larger, they get closer to 68%.

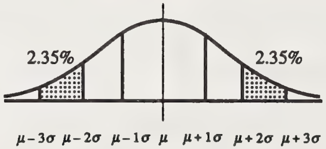
- The percentage of the total area (the percentage of the total value of data) within two standard deviations of the mean is 95%.
  There is 13.5% outside one standard deviation and inside two standard deviations on each side of the mean.



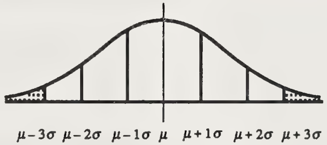


- The percentage of the total area (the percentage of the total value of data) within three standard deviations of the mean is 99.7%.
  There is 2.35% outside two standard deviations and inside three standard deviations on each side of the mean.

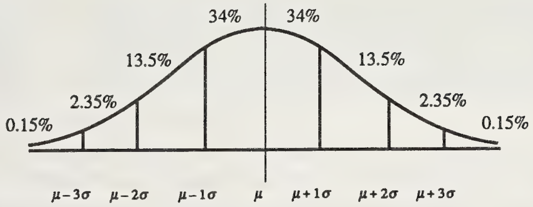




- The percentage of the total area (the percentage of the total value of data) outside three standard deviations of the mean is 0.3%. There is 0.15% on each side of the mean.



All of the previous information is given on this graph.

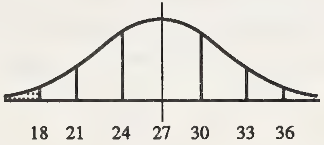


When solving problems that deal with the normal distribution, it is advisable to make a drawing of this graph and replace the bottom labels with the information from the problem.

Try using this idea with a problem.

A small electrical company is producing batteries that have an average life of 27 h and a standard deviation of 3 h. The company plans to replace any battery that does not last at least 18 h. If the company makes 200 000 batteries, how many batteries can it expect to replace? (Assume that this situation is normally distributed.)

Put this information on the normal distribution.



The area in which you are interested under the curve is to the left of the 18. This part of the graph is represented by 0.15%. Therefore, the company will replace 0.15% of the batteries that it produces.

0.15% of 200 000 is 0.15% × 200 000.

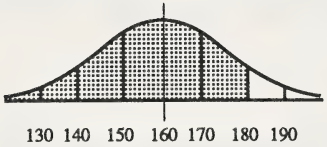
$$\begin{aligned} 0.15\% \times 200\,000 &= 0.0015 \times 200\,000 \\ &= 300 \end{aligned}$$

The company would expect to replace 300 batteries.

Now consider this problem.

A television station has decided to broadcast a local hockey game. From its statistics, it realized that the average game is 160 minutes long with a standard deviation of ten minutes. How much time should it schedule for the game to be 97% sure that it will broadcast the entire game without disrupting normally scheduled programming?

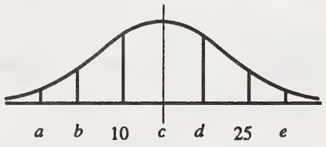
First draw a normal distribution and enter the given information.



From the graph, you can see that approximately 97% of all games are completed in less than two standard deviations above the mean. This means the station should allow 180 minutes (160 min + 2 × 10 min) or three hours.

Answer any three of the following questions. Assume a normal distribution exists.

1. Complete the following graph.



2. A small company packages chocolates. The mean mass of the chocolates is 45 g with a standard deviation of 2 g.

   a. If the company produces 10 000 chocolates in a production run, how many of the chocolates will be within 2 g of the required mass?

   b. How many chocolates will be rejected in a production run of 100 000 if the mass must be between 43 g and 49 g?

3. A large appliance company has decided to boost its sales by offering a replacement guarantee. However, it does not want to replace more than 2.5% of the appliances that it manufactures. If the average life of the appliance is fifteen years with a standard deviation of two years, for what length of time should the company make the guarantee?

4. What percent of the normally distributed population will lie between the following?

   a. μ and $\mu + \sigma$

   b. $\mu - 2\sigma$ and $\mu - 1\sigma$

   c. $\mu + 2\sigma$ and $\mu - 1\sigma$

   d. $\mu - 3\sigma$ and $\mu + 3\sigma$

✓ For solutions to **Activity 2**, turn to **Appendix A, Topic 1**.

You may study the next two activities by doing either **Part A** or **Part B** or both. **Part A** covers the objective with an audiotape, while **Part B** covers the objective through the print mode.

Whichever way you choose to study these activities, do the questions at the end of **Part B**.

Activity 3

Use z-scores to solve situations that are normally distributed.

Part A

Audio Activity

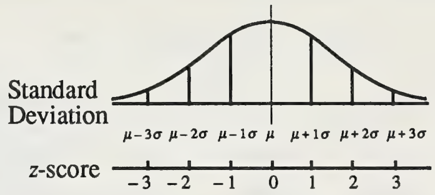
Insert the audiotape titled *Mathematics 30 Unit 6 – z-Scores* into your tape recorder and follow the instructions on the tape.

z-Scores

1

z-scores: units of standard deviation that allow fractional deviations from the mean

2 Comparison of z-scores and standard deviation



3

Formula for any z-score:

$$z=\frac{x-\mu}{\sigma}$$

4

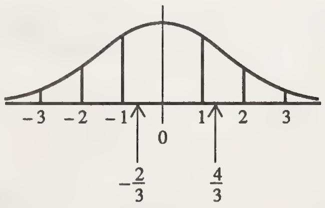
When $x = 13$,

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{13-15}{3} \\ &= \frac{-2}{3} \end{aligned}$$

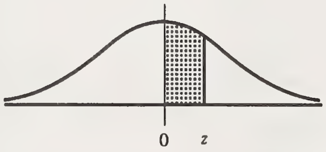
When $x = 19$,

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{19-15}{3} \\ &= \frac{4}{3} \end{aligned}$$

5



6



7

1.65 = 1.6 + 0.05

8

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

The area for a *z*-score of 1.65 is 0.4505. As a percentage, the area is 0.4505 × 100 or 45.05%.

9

z-score	area
– 1.41	0.4207
1.41	0.4207

10

z-score for $x = 50\%$

$$z = \frac{x-\mu}{\sigma}$$
$$= \frac{50-50}{12}$$
$$= \frac{0}{12}$$
$$= 0$$

z-score for $x = 75\%$

$$z = \frac{x-\mu}{\sigma}$$
$$= \frac{75-50}{12}$$
$$= \frac{25}{12}$$
$$= 2.08$$

The area for a *z*-score of 2.08 is 0.4812.

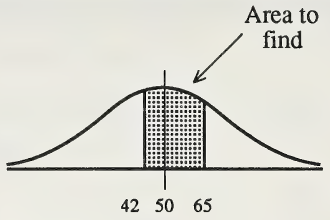
11

The number of students with a mark between 50% and 75% $= 0.4812 \times 80$

$= 38.496$

$\doteq 38$ (to the nearest whole number)

12



13

Area between 42% and 50%

$$z = \frac{42-50}{12}$$
$$= \frac{-8}{12}$$
$$= -0.67 \quad \text{(Ignore negative.)}$$

Area $= 0.2486$

Area between 50% and 65%

$$z = \frac{65-50}{12}$$
$$= \frac{15}{12}$$
$$= 1.25$$

Area $= 0.3944$

Total area $= 0.2486 + 0.3944$

$= 0.643$

14

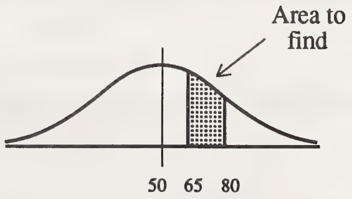
Number of students with marks between 42% and 65%

$= 0.643 \times 80$

$= 51.44$

$= 51$

15



16

Area between 50% and 80%

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{80-50}{12} \\ &= \frac{30}{12} \\ &= 2.50 \end{aligned}$$

Area = 0.4938

Area between 50% and 65%

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{65-50}{12} \\ &= \frac{15}{12} \\ &= 1.25 \end{aligned}$$

Area = 0.3944

The area between 65% and 80% is 0.4938 – 0.3944 = 0.0994.

17

Number of students who received marks between

$$\begin{aligned} 65\% \text{ and } 80\% &= 0.0994 \times 80 \\ &= 7.952 \\ &\doteq 8 \end{aligned}$$

Part B

In the last activity, you were able to compare data from different sets by finding the mean and the standard deviation and then comparing how the data was distributed according to the number of standard deviations.

Using the number of standard deviations gave you a convenient measure that was used with all types of data that were normally distributed.

Up to this point you are able to work only with whole units of standard deviations. A system needs to be developed that will allow you to work with fractional units of standard deviation.

The units that you are going to use are called *z*-scores. Standard deviation units are used with *z*-scores.

For example, note the following:

One standard deviation to the right of the mean will be a *z*-score of 1.
Two standard deviations to the right of the mean will be a *z*-score of 2.
Three standard deviations to the right of the mean will be a *z*-score of 3.

One standard deviation to the left of the mean will be a *z*-score of – 1.
Two standard deviations to the left of the mean will be a *z*-score of – 2.
Three standard deviations to the left of the mean will be a *z*-score of – 3.

As you can see, a sign is assigned to the *z*-score to represent the direction from the mean. Positive is to the right of the mean and negative is to the left of the mean.

The following formula will be used to find the *z*-score for any value in the set of data.

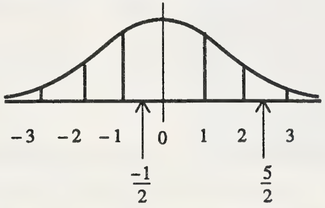
$$z\text{-score} = \frac{x-\mu}{\sigma}$$

In some situations, using the mean and the standard deviation with particular values of *x* will result in a fractional number for the *z*-score.

Start with two values, 11 and 17, from a set of data that has a mean of 12 and a standard deviation of 2. What are the *z*-scores for these two values?

$$\begin{aligned} z\text{-score} &= \frac{x-\mu}{\sigma} \\ &= \frac{11-12}{2} \\ &= \frac{-1}{2} \end{aligned} \qquad\qquad \begin{aligned} z\text{-score} &= \frac{x-\mu}{\sigma} \\ &= \frac{17-12}{2} \\ &= \frac{5}{2} \end{aligned}$$

Put these *z*-scores on the normal distribution.

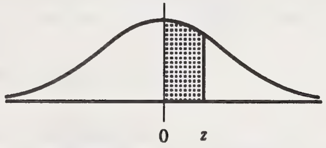


In the last activity, the standard deviation was an intermediate tool that was used to find the percentage of the area under the graph. What are the areas for the *z*-scores?

In **Appendix B** of this unit is a table labelled **The Standard Normal Distribution Table**. This is the table that will be used to find the areas.

Check to see how to use this table.

On the top of the table is a normal distribution graph like the one shown here.



The numbers given in the table represent the area between the 0 (or mean) and the *z*-score. The numbers that represent the area are written as decimal numbers from 0 to 0.4990, where 0 will represent no area under the graph and 0.4990 will represent almost half of the area under the graph.

But how do you use the table to find the areas?

This table is designed to give you the areas for *z*-scores that are given to the second decimal place. Take a look at how to use this type of table.

Find the area represented by the *z*-score 2.17. First break this number into two addends.

$$2.17 = 2.1 + 0.07$$

As you can see from this example, the first addend is the ones and tenths of the *z*-score and the second addend is the hundredths of the *z*-score.

The first addend is found in the far left column of the table. You will move horizontally from this point.

The second addend is found in the top row of the table. You will move down from this point.

The area that the two rectangles have in common will be the area for the *z*-score of 2.17, which is 0.4850.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890

This number can easily be converted into percentage form.

$$\begin{aligned} 0.4850 &= 0.4850 \times 100\% \\ &= 48.50\% \end{aligned}$$

The same table can be used for finding the areas for negative z-scores. Since the normal distribution is symmetrical about the 0 or the mean, the area for a negative z-score is calculated in the same way.

For example, the area for a z-score of – 1.17 is 0.3790, the same as for the z-score 1.17.

Try this out in a couple of different situations.

120 students took the test. How many students received a mark between 50% and 75% on a test that has a mean of 50% and a standard deviation of 15%?

First distinguish what z-scores you are looking for.

The mark of 50% is also the mean. This will be where you will start your measurement. (Remember, The Standard Normal Distribution Table always measures from the mean.)

You can ignore this mark for the time being.

The second mark, 75%, has the following z-score:

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{75-50}{15} \\ &= \frac{25}{15} \\ &= 1.67 \end{aligned}$$

This z-score, 1.67, is represented by the area 0.4525.

The product of this area and 120 (the total number of students) will give the number of students who scored between 50% and 75% on the test.

$$0.4525 \text{ of } 120 \text{ is } 0.4525 \times 120.$$

$$\begin{aligned} 0.4525 \times 120 &= 54.3 \\ &\doteq 54 \end{aligned}$$

54 of the students scored between 50% and 75% on the test.

How many students scored between 40% and 72% on the same test?

This time you are going to find the area between 50% and 72% and the area between 40% and 50%. The area that you are looking for will be the sum of these two areas.

The z-score for 50% can be calculated.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{50-50}{15} \\ &= \frac{0}{15} \\ &= 0 \end{aligned}$$

Determine the area between 50% and 72%.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{72-50}{15} \\ &= \frac{22}{15} \\ &= 1.47 \end{aligned}$$

The area for this *z*-score is 0.4292.

Now determine the area between 40% and 50%.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{40-50}{15} \\ &= \frac{-10}{15} \\ &= -0.67 \end{aligned}$$

The area for this *z*-score is 0.2486. (To look this *z*-score up in the table, ignore the negative sign.)

Now find the total area.

$$0.4292 + 0.2486 = 0.6778$$

Use the area to find the number of students.

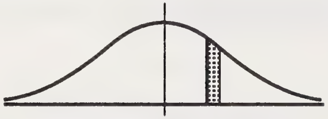
$$0.6778 \text{ of } 120 \text{ is } 0.6778 \times 120.$$

$$\begin{aligned} 0.6778 \times 120 &= 81.336 \\ &\doteq 81 \end{aligned}$$

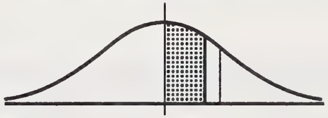
81 students received a mark between 40% and 72%.

How many students received a mark between 72% and 75%?

In this particular case, the area between 50% and 72% will not be part of the answer. This much area must be removed from the area for 50% to 75%. This means that the area for 72% must be subtracted from the area for 75%.



This is the area you are looking for.



Therefore, this area must be subtracted from the area of 50% to 75%.

$$\begin{aligned} &\text{Area of } 75\% - \text{Area of } 72\% \\ &= 0.4525 - 0.4292 \\ &= 0.0233 \end{aligned}$$

$$0.0233 \text{ of } 120 \text{ is } 0.0233 \times 120.$$

$$\begin{aligned} 0.0233 \times 120 &= 2.796 \\ &\doteq 3 \end{aligned}$$

Three students received a mark between 72% and 75%.

Remember: the table measures the area away from the mean. The mean here is 50%.

Locating the correct area under the normal distribution is crucial for solving the problem. If you want a more detailed explanation, turn to the **Extra Help** section at the end of this topic.

Do any three of the following questions.
Assume a normal distribution exists.

1. What is the *z*-score for the distance of two standard deviations left of the mean?

2. The heights for ten basketball players are given below. What are the *z*-scores for each height?

   188, 190, 192, 196, 196, 198, 199, 202, 202, and 207

3. Last year at the Acmy Golf and Country Club, golfers had a mean of 43 with a standard deviation of 2 on the front nine and a mean of 41 with a standard deviation of 1.8 on the back nine. Andrea shot a 40 on the front nine and a 38 on the back nine. On which half of the course did Andrea do better relative to the rest of the golfers?

4. During a major-league career of 25 seasons, a baseball player averaged 150 hits a season with a standard deviation of 20 hits. In how many seasons did the player have at least 185 hits?

5. The expected working years of a refrigerator are normally distributed with a mean of seven years and a standard deviation of one year. In one year, the manufacturer sold 1200 refrigerators. Determine the number of refrigerators working longer than eight years. The manufacturer will replace any refrigerator that stops working when it is less than five years old. Determine the number of refrigerators that must be replaced.

6. An appliance distribution company knows that the average life of a certain toaster is 9.6 years and the standard deviation is 5.5 years. The company does not want to replace more than 9% of the toasters that are sold. What guarantee to the nearest tenth of a year should the company provide in order to satisfy its replacement policy?

7. From past records, a manufacturing company knows that its air compressors have an average working period of twelve years with a standard deviation of three years. What percentage of these compressors will have to be repaired under warranty if the manufacturer guarantees these compressors for eight years? Assume a normal distribution and give the percentage to the nearest whole number.

For solutions to **Activity 3**, turn to **Appendix A, Topic 1**.

Activity 4

Use z-scores to calculate the probability of an event happening.

Part A

Audio Activity

Continue with the audiotape titled *Mathematics 30 Unit 6 – z-Scores*. When you have completed the tape, answer the questions at the end of **Part B** of this activity.

z-Score and Probability

1 Definition of Probability

Probability is a branch of mathematics concerned with the study of the chance that a given event will occur.

The probability of an event happening is the total number of favourable outcomes divided by the total number of outcomes where all of the outcomes have an equal chance of happening.

The probability of an event is a real number between 0 and 1 inclusive that indicates how likely it is that an event will happen.

2

outcomes: different results that can happen in an experiment

favourable outcomes: outcomes that are the ones you want

sample space: all outcomes taken together

event: all favourable outcomes

3

$$\text{Probability of an event} = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}$$

$$P(4) = \frac{1}{6}$$

4

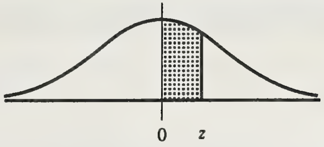
$$P(7) = \frac{0}{6}$$
$$= 0$$

Impossible event

$$P(<7) = \frac{6}{6}$$
$$= 1$$

Certain event

5



The entire area under the curve is equivalent to one.
Shaded area = 1σ = z-score of 1 = probability of 0.3413

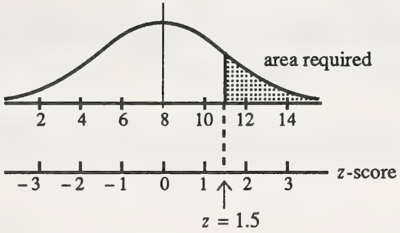
6

Mean $(\mu) = 8$ days, $\sigma = 2$ days
Probability cold lasts longer than 11 days = ?

$$z = \frac{x-\mu}{\sigma}$$
$$= \frac{11-8}{2}$$
$$= \frac{3}{2}$$
$$= 1.5$$

The area from the chart for $z = 1.5$ is 0.4332.

7



8

Area required = 0.500 − 0.4332
= 0.0668

The probability that a cold lasts more than eleven days is 0.0668.

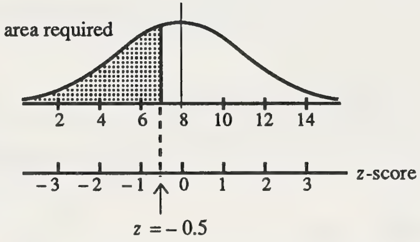
9

Mean $(\mu) = 8$ days, $\sigma = 2$ days
Probability you are rid of the cold in less than 7 days = ?

$$z = \frac{x-\mu}{\sigma}$$
$$= \frac{7-8}{2}$$
$$= \frac{-1}{2}$$
$$= -0.5$$

The area for $z = -0.5$ is 0.1915.

10



11

Area required = 0.5000 − 0.1915
= 0.3085

The probability of getting rid of a cold in less than seven days is 0.3085.

Part B

What is probability?

The following are different definitions that you may find for probability.

- Probability is a branch of mathematics concerned with the study of the chance that a given event will occur.
- The probability of an event happening is the total number of favourable outcomes divided by the total number of outcomes where all of the outcomes have an equal chance of happening.
- The probability of an event is a real number between 0 and 1 inclusive that indicates how likely it is that an event will happen.

These three definitions are completely different, and yet all three of them are correct. The definitions give you all of the information that you need to understand what probability is.

Take a look at a simple situation to help you understand this information.

The chance of rolling a five on a die is one out of six.

A die has six faces with dots that represent the numbers 1 through 6. Therefore, all of the possible outcomes of rolling a die are 1, 2, 3, 4, 5, and 6. All of the possible outcomes of the experiment are referred to as the **sample space**.

The **outcomes** are all of the different results that can happen. In this case, there are six different outcomes. All of the outcomes are the sample space and they all must be equally likely to happen.

The **event** is made up of all of the favourable outcomes. In this case, you are looking at the event that a five will be rolled on the die.

The **favourable** or **desirable outcomes** are the results that you want to appear. In this case, the number five is the only favourable outcome.

According to the second definition,

$$\text{probability of an event} = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}.$$

In this case, you would write $P(\textit{roll a five}) = \frac{1}{6}$.

Now you can use this equation to make a statement.

The chance of rolling a five on a die is one out of six.

As the first definition says, a probability statement is making a prediction about the future. If you rolled a die, there is one chance out of six that a five will come up, or if the die is rolled six times, a five will come up once.

The third definition puts a limit on the type of numbers that can be used for probability. As you have already seen, probability gives you a rational or fractional number. The last definition puts a lower limit of 0 and an upper limit of 1 on the number used for probability.

Make a closer examination of these two cases.

What is the chance of getting a seven on a die?

$$\text{Probability of an event} = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}$$

$$\begin{aligned} P(roll\ a\ seven) &= \frac{0}{6} \\ &= 0 \end{aligned}$$

This case examines the situation where there is no chance of the event happening. This situation is referred to as an **impossible event**. There is no way you can get a smaller probability than 0 from this formula.

What is the chance of getting a natural number that is less than seven on the roll of a die?

$$\text{Probability of an event} = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}$$

$$\begin{aligned} P(<seven) &= \frac{6}{6} \\ &= 1 \end{aligned}$$

This case examines the situation where the event must happen. This situation is referred to as a **certain event**. Since all of the outcomes are also the favourable outcomes, it is not possible to get a probability higher than 1.

Another method for finding the probability of a situation is to use z-scores.

The areas in **The Standard Normal Distribution Table** are already set up to represent probabilities. The table considers the entire area under the graph to be 1. Each of the areas in the table will then represent the probability of an event happening that is between the mean and the stated z-score.

See how this is done through an example.

Example 1

The mean duration of a common cold is ten days with a standard deviation of two days. What is the probability that a cold will last longer than thirteen days?

Solution:

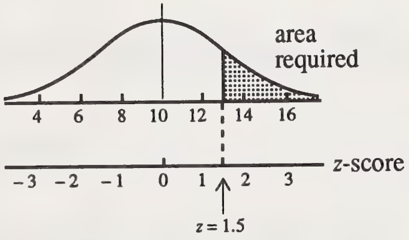
First use the z-score formula.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{13-10}{2} \\ &= 1.5 \end{aligned}$$

The area associated with a z-score of 1.50 is 0.4332.

This area is the amount between the mean and the *z*-score. According to the question, you are looking for the area associated with the *z*-score. Therefore, to find this required area, the area found in the table must be subtracted from 0.5 (the area for half of the graph).

$0.5 - 0.4332 = 0.0668$



The probability of having a common cold longer than thirteen days is 0.0668.

Do any three of the following questions.

1. Find the probability of getting the following *z*-scores.

   a. $P(z<0)$

   b. $P(0<z<2)$

   c. $P(z>1.80)$

   d. $P(-0.89<z<1.24)$

   e. $P(-2.37<z<-0.92)$

2. The average life expectancy of a cat is thirteen years with a standard deviation of 1.3 years. What is the probability that a cat will live for the following periods of time?

   a. at least fifteen years

   b. at least twelve years

   c. less than ten years

   d. less than fourteen years

3. A firm manufactures batteries that have a mean life of 35 h and a standard deviation of 6.5 h. What is the probability that a battery will last for the following periods of time?

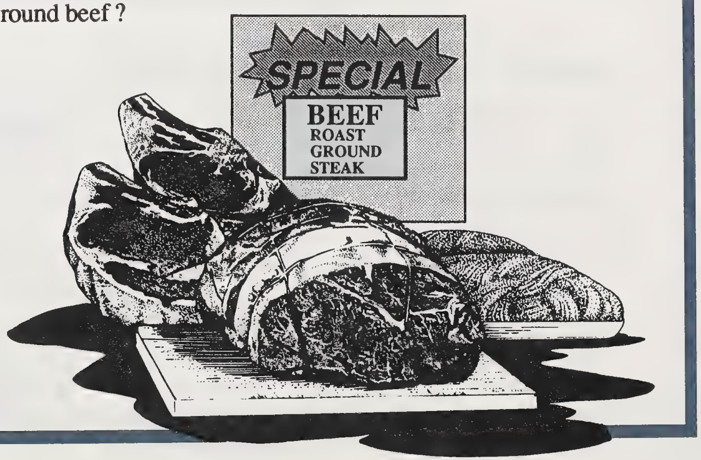
   a. less than 32 h

   b. less than 43 h

   c. more than 47 h

   d. more than 33 h

4. A supermarket leaves packaged ground beef on the shelf for an average of three days. If there is a probability of 0.4826 of selecting a package which has been sitting on the shelf for 1.3 days, what is the standard deviation of the shelf period of the ground beef ?



5. Use the following table to determine the probability, to two decimal places, that a fluorescent light tube will last longer than 820.5 h.

Lifetime of 200 Fluorescent Tubes	
Lifetime (h)	Frequency
600.5 - 650.5	12
650.5 - 700.5	18
700.5 - 750.5	20
750.5 - 800.5	39
800.5 - 850.5	68
850.5 - 900.5	27
900.5 - 950.5	16

6. A manufacturer of thumbtacks determines the mean number is 103 and the standard deviation is 2 for a box of thumbtacks advertised as containing 100 tacks. Determine the probability that a box will contain 100 or more tacks. Assume a normal distribution.

✓ For solutions to **Activity 4**, turn to **Appendix A, Topic 1.**

If you require help, do the Extra Help section.

If you want more challenging explorations, do the Extensions section.

} You may decide to do both.

Extra Help

Part A

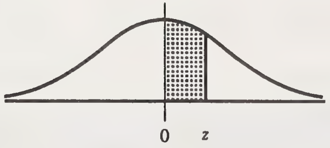
If you have access to a videocassette player, you may find it helpful to view the program "Graphing Statistics". This is the last program on the videocassette titled *Graphing Mathematical Concepts*.[1] Pay particular attention to the section on *z*-scores and area as well as the section on *z*-scores and probability.

Part B

Selecting the correct area under the normal distribution graph is crucial to solving problems. In this section you are going to take a closer look at how to select the appropriate area(s) under the graph and the procedure that must be followed to solve for that area.

The Standard Normal Distribution Table is given in **Appendix B**.

The Standard Normal Distribution Table always measures the area away from the mean or the 0 *z*-score to a positive *z*-score. On the normal distribution, this is always to the right of the 0 *z*-score, as shown in the following graph.



Since the table will only give areas to the right of the *z*-score, the largest possible area that you can have is 0.5 or half of the area under the graph.

Take a minute to examine the chart. The largest area given in the chart is 0.4990 or almost half of the graph. That leaves only an extremely small area to the right of the 3.09 *z*-score line. This area is so small that it can usually be omitted without influencing your conclusions.

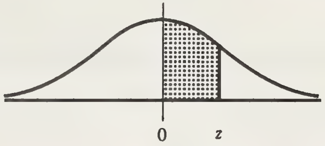
Also notice that the last two areas in the table are the same. The area at the extreme ends of the graph will only increase by an extremely small percentage for each additional *z*-score.

[1] *Graphing Mathematical Concepts* is a title of Access Network.

Take a look at how to find each of the types of areas using the graph and the table.

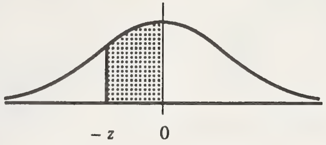
Type 1: 0 *z*-Score to Positive *z*-Score

The easiest area to find is the area from the mean (or the 0 *z*-score) to a positive *z*-score. Since the table is already set up for finding this type of area, you can just read the area from the table.



Type 2: 0 *z*-Score to Negative *z*-Score

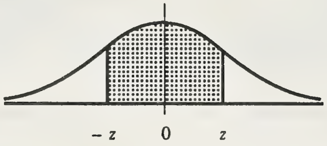
The next type of area is from the mean to a negative *z*-score. Since the normal distribution is symmetrical, a specified distance left of the mean will yield the same area as that distance right of the mean.



Here all you need to do is to ignore the negative sign on the *z*-score and look up the area in the chart as if you were looking for the area for a positive *z*-score.

Type 3: From Negative to Positive *z*-Score

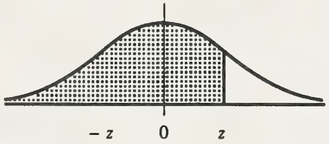
This type of area is broken up into two different areas: an area from the 0 *z*-score to the positive *z*-score and an area from the 0 *z*-score to the negative *z*-score.



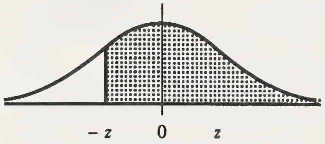
Now you are looking for Type 1 and Type 2 areas.

The total area is then found by adding the two individual areas.

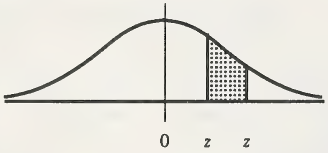
Also Type 3



Also Type 3



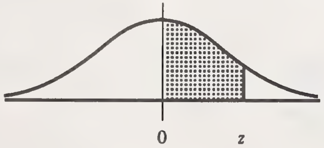
Type 4: From Positive to Positive *z*-Score



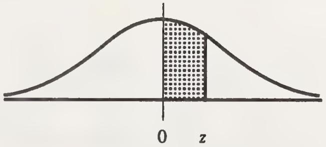
This type of area is also broken up into two different areas. The first area will be from the 0 z-score to the higher positive z-score and the second area will be from the 0 z-score to the lower positive z-score. The two areas that you will be dealing with are shown.

A Type 4 area is then found by taking the difference of the two areas.

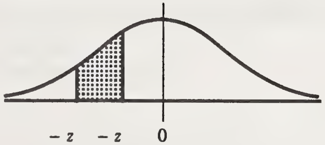
Take this area.



Then subtract this area.



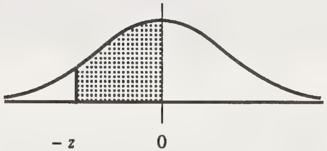
Type 5: From Negative to Negative z-Score



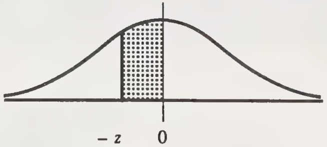
This type of area is found using the same procedure that was used for the Type 4 area.

The area is broken up into two areas. The first area is from the 0 z-score to the smaller negative z-score and the second area is from the 0 z-score to the larger negative z-score.

Take this area.



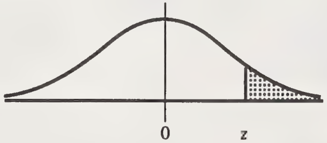
Then subtract this area.



The difference of the two areas is then found.

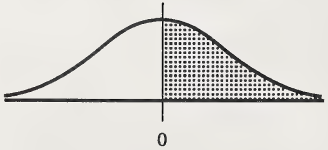
Type 6: Positive Extreme Area

This type of area is on the right side of a positive z-score.

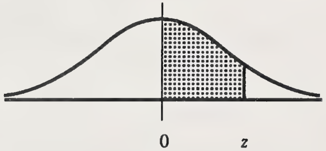


To find this area, first find the Type 1 area from the 0 z-score to the positive z-score.
Since half of the graph has an area of 0.5, the Type 1 area is subtracted from 0.5.

Take this area.



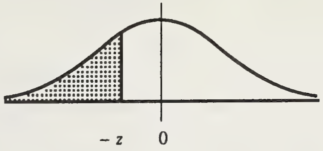
Then subtract this area.



What is left is the extreme positive area of a Type 6.

Type 7: Negative Extreme Area

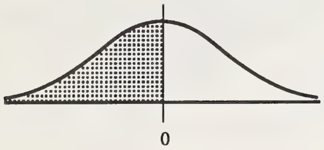
This type of area is on the left side of a negative *z*-score.



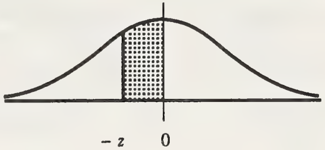
To find this area, first find the Type 2 area from the 0 *z*-score to the negative *z*-score.

Since half of the graph has an area of 0.5, the Type 2 area is subtracted from 0.5.

Take this area.



Then subtract this area.



What is left is the extreme negative area of a Type 7.

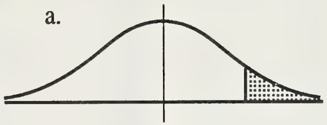
All other types of graphs are made up of combinations of these seven types. In those cases, you will break the graph down into the individual types and then find the areas.

You are now ready to try some questions.

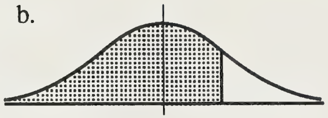
Do any three of the following questions.

1. Identify the type or types for each of the following.

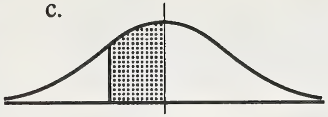
a.



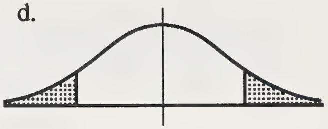
b.



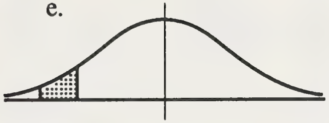
c.



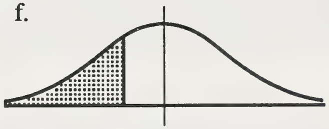
d.



e.



f.



2. Identify the type or types for each of the following. It may help to first make the graph.

a. $0 < z < 2.76$

b. $-2.82 < z < -1.09$

c. $z < -2.3$

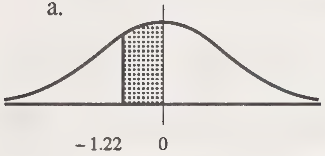
d. $z < -2$ and $z > 2$

e. $2.3 < z < 2.32$

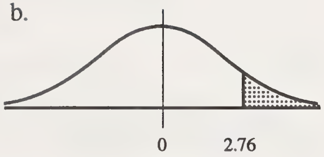
f. $-0.04 < z < 1.03$

3. Find the areas for the following.

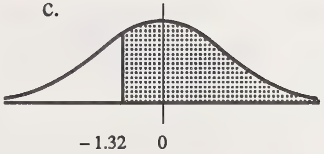
a.



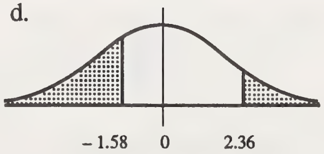
b.



c.



d.



4. Find the following areas.

a. $z < -1.02$

b. $z > -2.04$

c. $0 < z < 2.45$

d. $-2.38 < z < -1.89$

For solutions to **Extra Help**, turn to **Appendix A, Topic 1**.

Extensions

In this section you will be shown how to use a calculator with statistical functions in order to find the standard deviation of a set of grouped data.

You should refer to the operator's manual for your calculator to learn the proper way to use your calculator.

You will use the same data that was used earlier in the topic. In the following table, the frequency represents the number of days that a certain range of requests for 180 cm skis were made. The class limits state the range of the number of requests on any particular day.

Length: 180 cm

Class	Class Limits	Frequency	Class Midpoint
1	50 - 59	22	54.5
2	60 - 69	34	64.5
3	70 - 79	41	74.5
4	80 - 89	54	84.5
5	90 - 99	43	94.5
6	100 - 109	38	104.5
7	110 - 119	18	114.5

Step 1: Turn on the statistical function of the calculator. Press the $\boxed{\text{MODE}}$ key followed by the $\boxed{\cdot}$ key.

Step 2: Turn on the data receive mode of the calculator. This will tell the calculator that you are going to enter the data. Press the $\boxed{\text{INV}}$ key followed by the $\boxed{\text{SAC}}$ key.

Step 3: Enter the data.

The following procedure is used for entering the data. Enter the first midpoint. Press the $\boxed{\times}$ key. Enter the frequency for that class. Press the $\boxed{x}$ key. Repeat the procedure for the next midpoint.

For example:

54.5 $\boxed{\times}$ 22 $\boxed{x}$

64.5 $\boxed{\times}$ 34 $\boxed{x}$

74.5 $\boxed{\times}$ 41 $\boxed{x}$

84.5 $\boxed{\times}$ 54 $\boxed{x}$

94.5 $\boxed{\times}$ 43 $\boxed{x}$

104.5 $\boxed{\times}$ 38 $\boxed{x}$

114.5 $\boxed{\times}$ 18 $\boxed{x}$

Step 4: Find the standard deviation. To find the standard deviation, press the $\boxed{\text{INV}}$ key followed by the $\boxed{\sigma n}$ key. The calculator will then display the standard deviation for the data. In this case, it is 17.1112127 which is rounded to 17.1.

The mean can be found by pressing the $\boxed{\text{INV}}$ key followed by the $\boxed{\bar{x}}$ key. This gives 84.42 which agrees with the value obtained earlier in the topic.

The $\boxed{\sigma n}$ key is the $\boxed{8}$ key.

The $\boxed{x}$ key is the $\boxed{\text{M+}}$ key.

The $\boxed{\bar{x}}$ key is the $\boxed{7}$ key.

The number of values can be found by pressing the INV key followed by the n key. This gives 250 which is the frequency for this length of skis.

Answer one of the following questions.

1. Find the mean and the standard deviation for the following grouped data.

Class	Class Limits	Frequency
1	12 - 20	4
2	21 - 29	12
3	30 - 38	32
4	39 - 47	76
5	48 - 56	59
6	57 - 65	18
7	66 - 74	11
8	75 - 83	3

2. Find the mean and the standard deviation for the following grouped data.

Class	Class Limits	Frequency
1	22 - 26	8
2	27 - 31	10
3	32 - 36	29
4	37 - 41	57
5	42 - 46	73
6	47 - 51	43
7	52 - 56	22
8	57 - 61	9

✓ For solutions to **Extensions**, turn to **Appendix A, Topic 1**.

The n key is the 6 key.

Topic 2 Bivariate Data

Introduction

Usually you are interested in how a certain characteristic influences the outcome of a situation.

How does the speed you drive influence your gas mileage?

How does the diameter of a cable influence the amount of mass it can support?

How does the speed of a running back in football influence the number of yards rushed in a season?

Each of the above situations is made up of data that have two numbers. In the first case, for every speed that you test, you will get a specified gas mileage. These two numbers (speed, gas mileage) can be used to plot a single point on a graph. If enough points are plotted on a graph, a pattern may become evident.

The pattern is what you are going to be looking for in this topic.

What Lies Ahead

Throughout this topic you will learn to

1. plot sets of bivariate data to produce a scatterplot
2. plot a line of best fit on a scatterplot using the median fit method
3. use the equation of the line of best fit to generate new data for a population
4. determine the strength and type of correlation between the variables of a bivariate distribution
5. collect, organize, and analyze sets of bivariate data

Now that you know what to expect, turn the page to begin your study of bivariate data.

Exploring Topic 2

Activity 1

Plot sets of bivariate data to produce a scatterplot.

In studying statistics to this point, you have studied data that involved only one variable. However, it is frequently more valuable to study data that involves two variables. That is what bivariate data is: data that consists of two variables.

When bivariate data is plotted on a scatterplot, patterns may become evident that are not visible when examining the data alone.

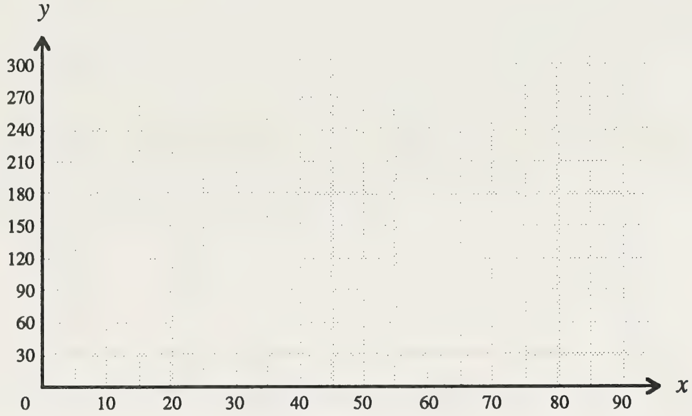
Use an example to plot and analyze a scatterplot.

The following chart contains the number of strikeouts and the number of bases on balls allowed by a number of pitchers for the 1989 season in the Distance Baseball League.

Pitcher	Bases on Balls	Strikeouts
Maddon, Calgary	81	140
Prune, Edmonton	44	131
Conic, Lethbridge	80	213
Daze, Medicine Hat	86	298
Garney, Grande Prairie	89	162
Drysday, High Level	50	127
Shoulder, Lloydminster	53	144
Jakson, Drumheller	71	161
Reussel, Red Deer	42	92
Hornblower, Jasper	73	178
Scottie, Banff	54	190
Suttuer, Calgary	70	144
Martten, Edmonton	55	120
Coodie, Lethbridge	57	175
Terence, Medicine Hat	34	65
Carrie, Grande Prairie	68	116
Smilie, Stettler	46	129
Mussen, St. Paul	58	112
Browner, St. Albert	64	124
Drowns, Wetaskiwin	47	118
Larry, Banff	56	180
Knocks, Red Deer	67	103

Step 1: Decide on which axis each variable will be placed. Decide on the scale for each axis.

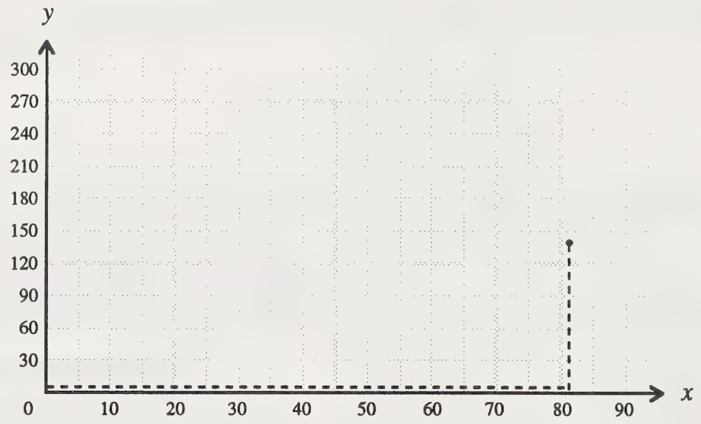
For this data, the number of bases on balls allowed will be placed along the horizontal axis and the number of strikeouts will be placed along the vertical axis. While examining the data, you will notice that the greatest number of bases on balls allowed by a pitcher was 89 and the greatest number of strikeouts was 298. You will have to make sure that both of these figures will fit onto the graph. Therefore, the horizontal axis will go to 90 and the vertical axis will go to 300. The graph used for the scatterplot is shown.



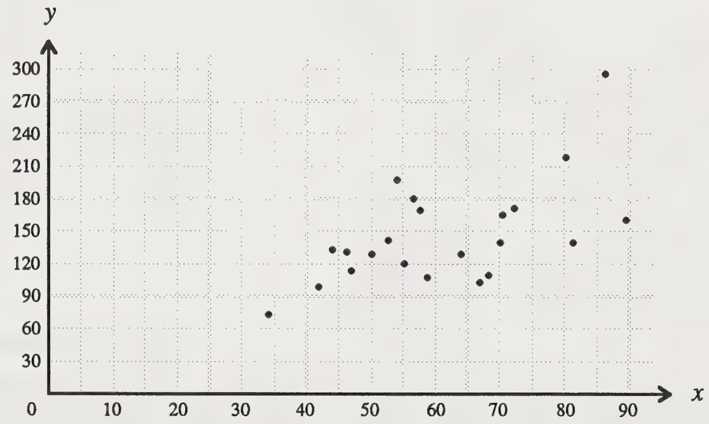
Step 2: Plot the data onto the scatterplot.

The data is plotted in exactly the same way you plot ordered pairs. The bases on balls allowed will be treated as the x-coordinate and the strikeouts will be treated as the y-coordinate.

To graph the point that represents Maddon, move from the origin 81 to the right and 140 up. This is the place where you will plot the first point. This movement is shown on the following graph.



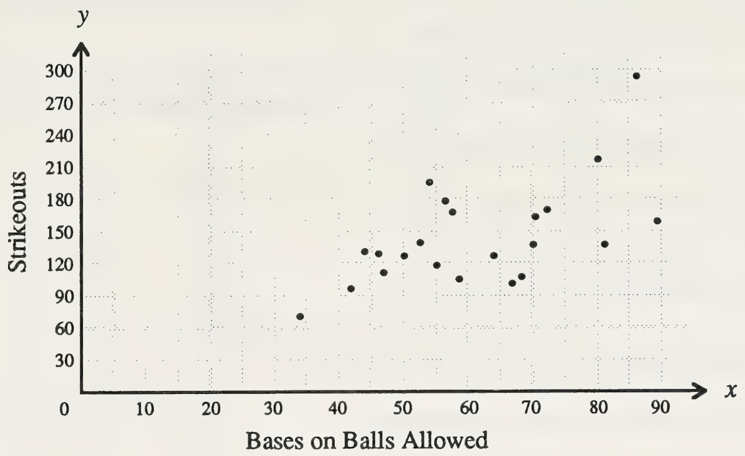
The rest of the points are plotted in the same manner. The scatterplot with all of the points plotted is shown.



Step 3: Label the scatterplot and the axes.

The labels have been placed on the following scatterplot.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



The scatterplot is now complete.

While examining this graph, you will notice a trend. The more bases on balls allowed by a pitcher, the more likely it is that the pitcher is going to have more strikeouts. You should note that an increase in one of these two characteristics will not cause an increase in the other characteristic. An increase in both of these characteristics will depend on an increase in a third characteristic, namely, how many batters the pitcher has faced.

Since these two characteristics do not depend on each other, it did not matter on which axis each characteristic was placed. It would have been just as correct to place the bases on balls allowed along the vertical axis and the number of strikeouts along the horizontal axis.

However, there are times when the two characteristics must be placed along a certain axis.

Whenever one characteristic is dependent upon the other characteristic, the following rule must be observed.

The independent characteristic is always placed along the horizontal axis (or the *x*-axis), and the dependent characteristic is always placed along the vertical axis (or the *y*-axis).

The following example looks at a case where there are independent and dependent characteristics.

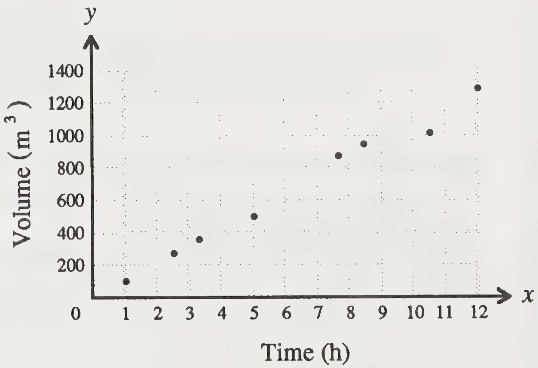
The following chart shows the amount of gas that has been transported through a pipeline after different lengths of time.

Time (h)	Volume (m^3)	Time (h)	Volume (m^3)
1	110	2.5	273
5	498	3.25	354
10.5	1056	8.4	915
12	1280	7.75	846

In this particular case, the volume that has been transported through the pipeline is dependent upon the amount of time the gas has been flowing.

Therefore, time (independent characteristic) will be placed along the horizontal axis and the volume (dependent characteristic) will be placed along the vertical axis.

Volume of Gas Transported Through Pipeline Omega
(January 1989)



Now you are ready to try some questions.

Do any three of the following questions.

1. For each of the following, identify whether the characteristics are independent, dependent, or have no relation.

   a. the height of people and the city in which they live

   b. the mass and volume of water in a beaker

   c. the number of goals and assists scored by hockey players

   d. the distance travelled after a given amount of time

2. Make a scatterplot for the following information.

The Number of Wins and the Earned Run Average of Starting Pitchers in the Distance Baseball League, 1989

Pitcher	Wins	ERA
Maddon, Calgary	18	3.18
Prune, Edmonton	12	2.44
Conic, Lethbridge	20	2.22
Daze, Medicine Hat	13	3.67
Garney, Grande Prairie	12	3.69
Drysday, High Level	15	3.08
Shoulder, Lloydminster	16	3.26
Jakson, Drumheller	23	2.73
Reussel, Red Deer	19	3.12
Hornblower, Jasper	23	2.26
Scottie, Banff	14	2.92
Suttuer, Calgary	13	3.86
Martten, Edmonton	15	2.72
Coodie, Lethbridge	18	3.19
Terence, Medicine Hat	9	2.92
Carrie, Grande Prairie	10	4.29
Smilie, Stettler	13	3.25
Mussen, St. Paul	16	3.43
Browner, St. Albert	18	3.41
Drowns, Wetaskiwin	13	3.32
Larry, Banff	17	2.91
Knocks, Red Deer	14	3.14

3. Make a scatterplot for the following information.

The Mass and Fuel Consumption of Vehicles

Mass (kg)	Fuel Consumption (L/100 km)	Mass (kg)	Fuel Consumption (L/100 km)
800	5.0	1100	7.5
1300	7.8	1800	12.3
900	6.2	1400	10.4
1100	8.2	1700	12.0
1200	9.6	1000	8.5
600	3.9	1200	8.2
1500	9.8	900	5.6

4. Make a scatterplot for the following information.

Federal and Provincial Income Tax Payable for Northland Before Surtax, 1989

Federal	Provincial	Federal	Provincial
820	381.20	1060	493.60
1540	716.40	1780	828.80
2040	949.90	2280	1060.30
2560	1190.10	2620	1218.20
2740	1274.40	2860	1330.60
2980	1386.80	3020	1404.20
3160	1469.10	3340	1553.40
3660	1702.60	4240	1972.90

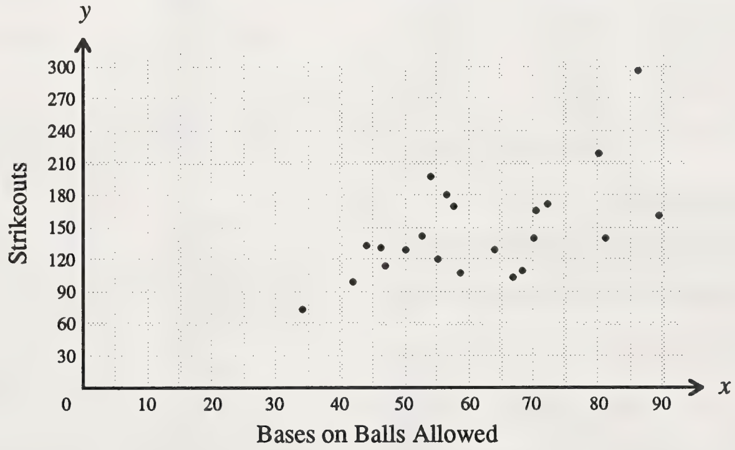
✓ For solutions to **Activity 1**, turn to **Appendix A, Topic 2**.

Activity 2

Plot a line of best fit on a scatterplot using the median fit method.

Take another look at the pitcher scatterplot that was drawn in the first activity.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



It is clear from the scatterplot that as the bases on balls allowed increase, so do the strikeouts. Unfortunately, the points on the scatterplot do not form a straight line.

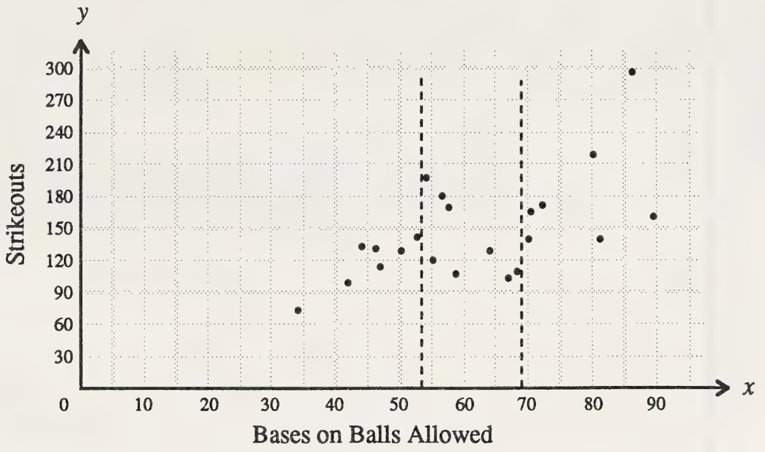
An approximation of the points can be formed to make a straight line. This approximation is referred to as **the line of best fit**.

The **median fit method** will be used to find the line of best fit.

Use this example to find the line of best fit using the median fit method.

Step 1: The points on the scatterplot are grouped into three vertical strips that have the same number of points. (If it is not possible for all three strips to have the same number of points, make sure the two outside strips have the same number of points.)

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



In this particular case, there are twenty-two points. The three strips are divided so that there are seven points in the first and last strips and eight points in the middle strip.

Step 2: The median point is found for each strip.

The first strip has the following points:
(34, 65), (42, 92), (44, 131), (46, 129), (47, 118), (50, 127), and (53, 144)

The median *x*-coordinate is as follows:
34, 42, 44, 46, 47, 50, 53

$\rightarrow 46$

The median *y*-coordinate is as follows:
65, 92, 118, 127, 129, 131, 144

$\rightarrow 127$

The second strip has the following points:
(54, 190), (55, 120), (56, 180), (57, 175), (58, 112), (64, 124), (67, 103), and (68, 116)

The median *x*-coordinate is as follows:
$54, 55, 56, \underline{57, 58}, 64, 67, 68$

$$\rightarrow \frac{57+58}{2}=57.5$$

The median *y*-coordinate is as follows:
$103, 112, 116, \underline{120, 124}, 175, 180, 190$

$$\rightarrow \frac{120+124}{2}=122$$

The median is the middle value from the data when the data is arranged in ascending order.

The median of a group of data that has an even number of values is the average of the two middle values.

The third strip has the following points:
(70, 144), (71, 161), (73, 178), (80, 213), (81, 140), (86, 298), and (89, 162)

The median x-coordinate is as follows:
70, 71, 73, 80, 81, 86, 89

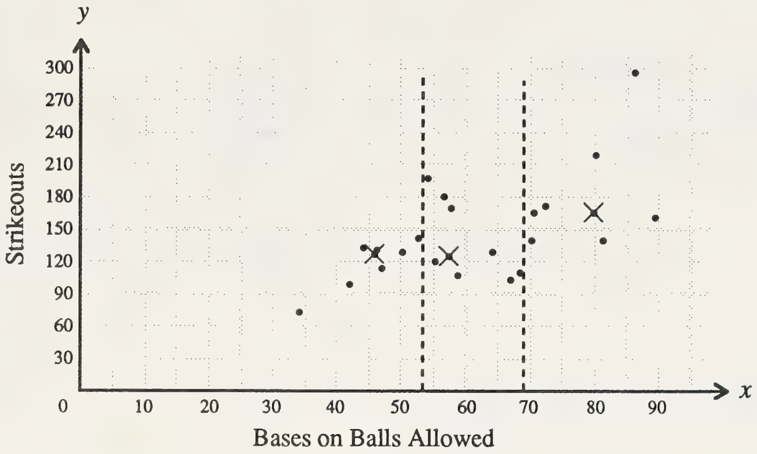
→ 80

The median y-coordinate is as follows:
140, 144, 161, 162, 178, 213, 298

→ 162

Step 3: These points are then plotted onto the scatterplot.

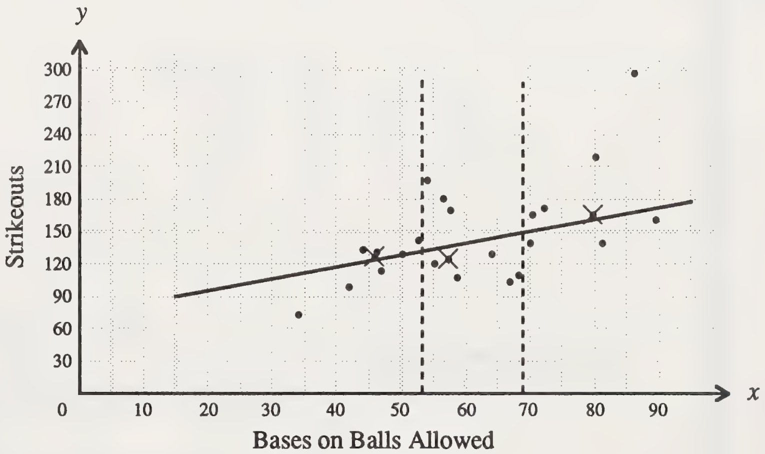
Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



The median point can also be found by visual inspection. See the **Extra Help** section at the end of the topic.

Step 4: Place a ruler on the median points of the two outside strips. Slide the ruler parallel to itself, one-third of the distance to the median point in the middle strip. Draw a straight line. This is the line of best fit.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989

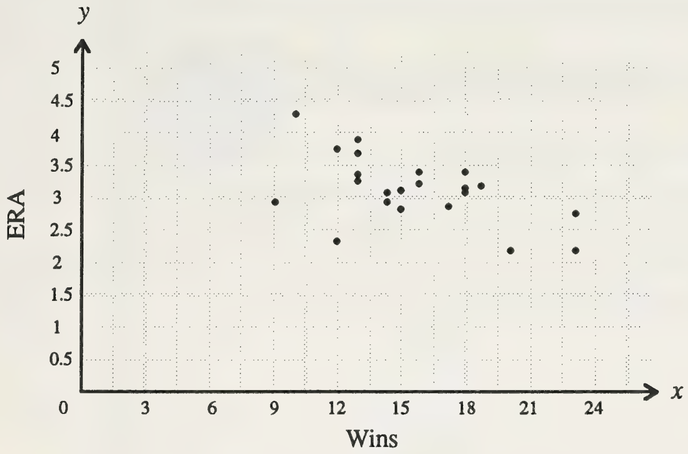


This line represents an approximation of the relation between the number of bases on balls allowed and the number of strikeouts by pitchers in the Distance Baseball League.

Do any two of the following questions.

1. Draw the line of best fit for the following scatterplot.

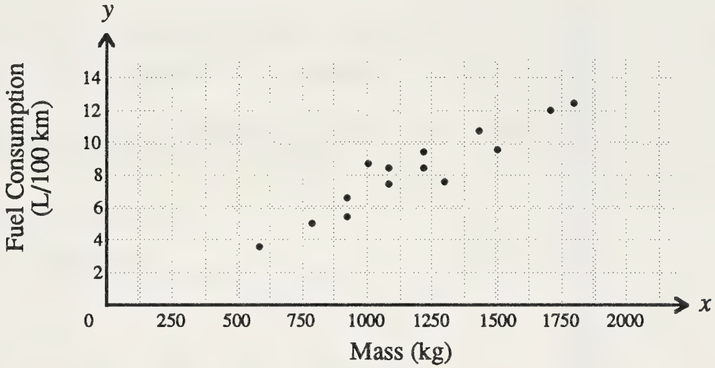
Wins versus ERA for Distance League Starting Pitchers, 1989



The coordinates of the points are given in Activity 1, question 2.

2. Draw the line of best fit for the following scatterplot.

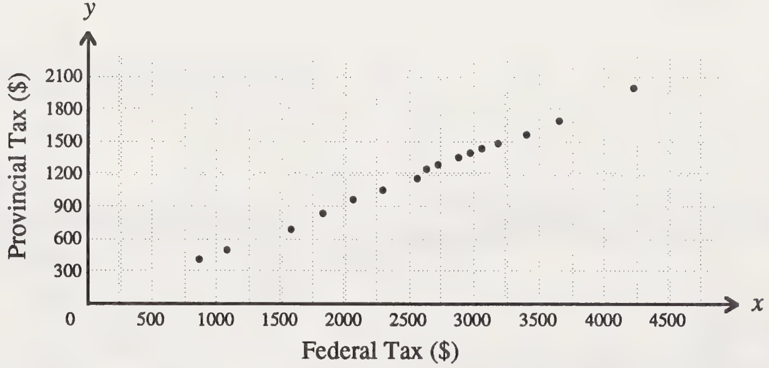
Mass versus Fuel Consumption of Vehicles



The coordinates of the points are given in Activity 1, question 3.

3. Draw the line of best fit for the following scatterplot.

Federal and Provincial Income Tax Payable in Northland
Before Surtax, 1989



The coordinates of the points are given in Activity 1, question 4.

✓ For solutions to **Activity 2**, turn to **Appendix A, Topic 2**.

Activity 3

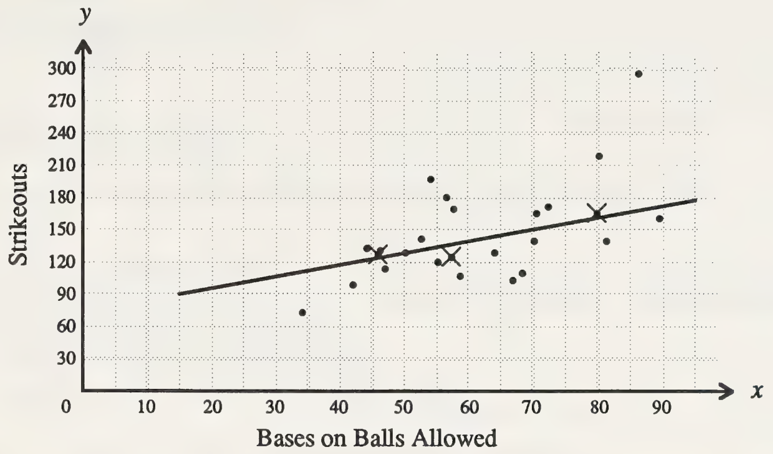
Use the equation of the line of best fit to generate new data for a population.

At this point you are able to take data, plot it on a scatterplot, and draw its line of best fit. Now you are ready to extend the width of the sample and find possible new values of data for the situation.

The key for extending the range of the data lies in the line of best fit. It is along this line that you will expect to find new pieces of data.

Take another look at the scatterplot that was made in the previous activity.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989

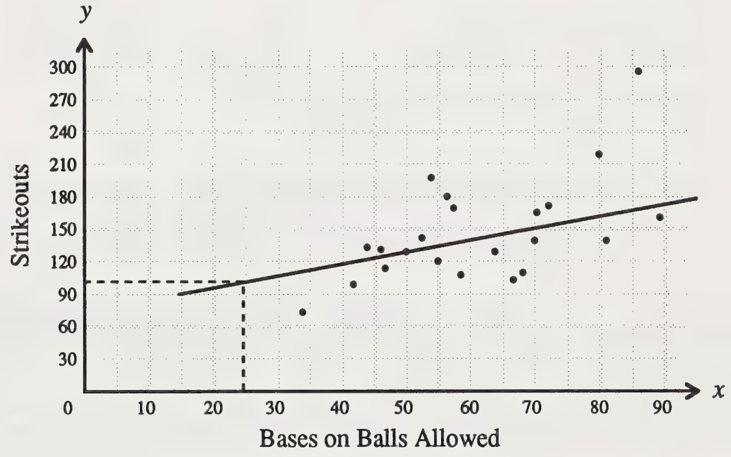


If a pitcher allowed twenty-five bases on balls, how many strikeouts would you expect the pitcher to have?

To find the solution to this question, you would go through the following steps:

1. Find 25 along the horizontal axis (or *x*-axis).
2. Move straight up to the line of best fit.
3. Move horizontally to the left scale.
4. Read the value from the left scale. This will be the number of strikeouts that you would expect the pitcher to obtain.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



The above scatterplot shows that the pitcher would get approximately 100 strikeouts.

What happens when one of the values is not on the graph? Take a look at a particular case in which this happens.

How many bases on balls would you expect a pitcher to allow if that pitcher threw 240 strikeouts?

In this case, you would have to start along the vertical or *y*-axis and move horizontally to the line of best fit. This is not possible with this scatterplot.

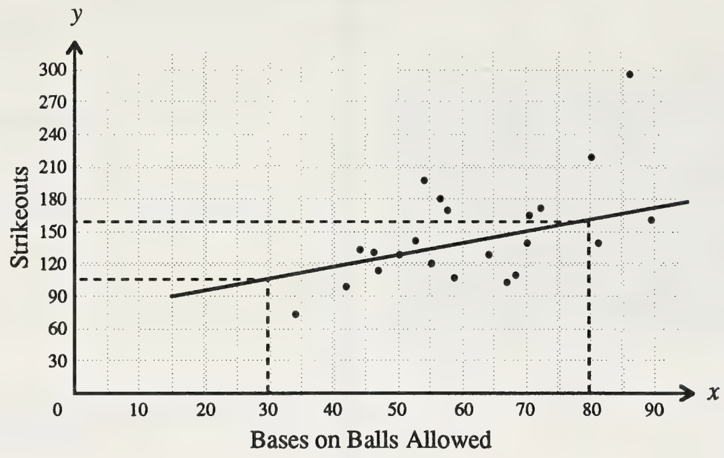
What you are looking for is a point along the line of best fit that has a *y*-coordinate of 240. This cannot be done visually.

If you find the equation of the line, you can solve for the missing coordinate algebraically.

Step 1: Find the equation of the line of best fit.

- First, find the coordinates of two points that are on the line.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



The two points selected are (30, 105) and (80, 155).

- Next, find the slope of the line.

$$\begin{aligned} m &= \frac{y_1 - y_2}{x_1 - x_2} \\ &= \frac{155 - 105}{80 - 30} \\ &= \frac{50}{50} \text{ or } 1 \end{aligned}$$

The slope of the line is 1.

- Then, solve for the y-intercept using the y-intercept form of the equation.

$$y = mx + b$$
$$y = 1x + b \text{ or}$$
$$y = x + b$$

Substitute the point (30, 105) into the equation and find the y-intercept.

$$y = x + b$$
$$105 = 30 + b$$
$$75 = b$$

Substituting the y-intercept into the equation will give you the equation of the line of best fit.

$$y = x + b$$
$$y = x + 75$$

The equation of the line of best fit is $y = x + 75$.

Step 2: Substitute the known coordinate into the equation of the line of best fit and solve for the missing coordinate.

$$y = x + 75$$
$$240 = x + 75$$
$$165 = x$$

This coordinate approximately represents the number of bases on balls allowed by a pitcher who throws 240 strikeouts.

This method will allow you to find new data (ordered pairs) anywhere along the line of best fit.

The point slope form can also be used to find the equation of the line.

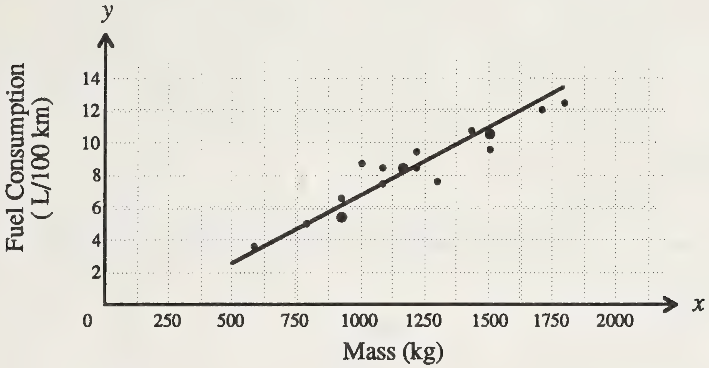
$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$
$$y - 105 = \frac{155 - 105}{80 - 30}(x - 30)$$
$$y - 105 = \frac{50}{50}(x - 30)$$
$$y - 105 = x - 30$$
$$y = x + 75$$

This method is more accurate than estimating the missing coordinate from a scatterplot. Use this method whenever a more accurate answer is desired or when you are given two ordered pairs on the line of best fit.

Answer any two of the following questions.

1. Find the following missing values using the given scatterplot.

Mass versus Fuel Consumption of Vehicles



a. What is the fuel consumption for vehicles that have the following masses?

i. 750 kg

ii. 1275 kg

iii. 1750 kg

iv. 400 kg

v. 4500 kg

b. What is the mass of each vehicle given the following amounts of fuel consumption?

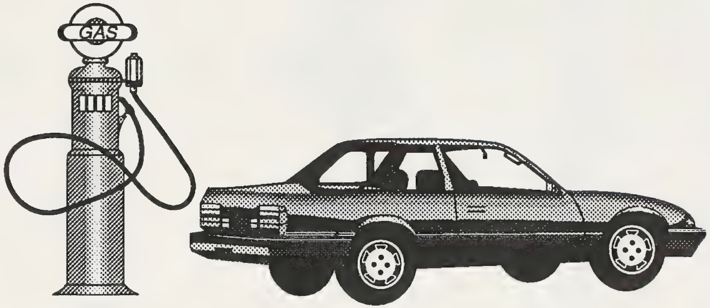
i. 4.2

ii. 8.8

iii. 12.5

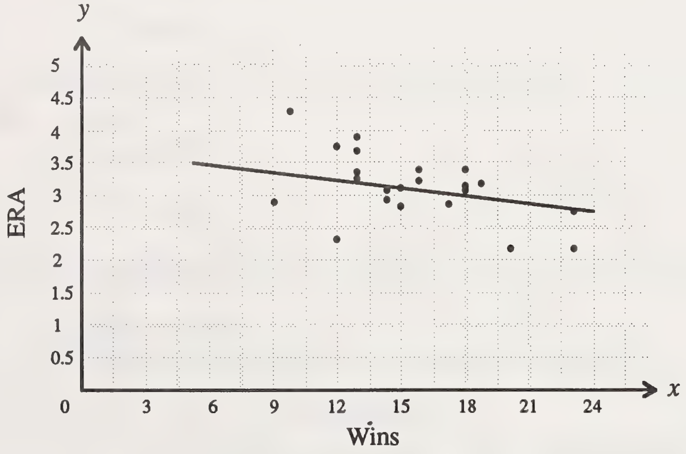
iv. 1.2

v. 17.9



2. Find the following missing values using the scatterplot.

Wins versus ERA for Distance League Starting Pitchers, 1989



a. Find the number of wins a pitcher will have if the pitcher has an ERA of the following:

i. 3.40

ii. 3.20

iii. 3.00

iv. 3.60

v. 1.90

b. Find a pitcher's ERA if the pitcher has the following numbers of wins:

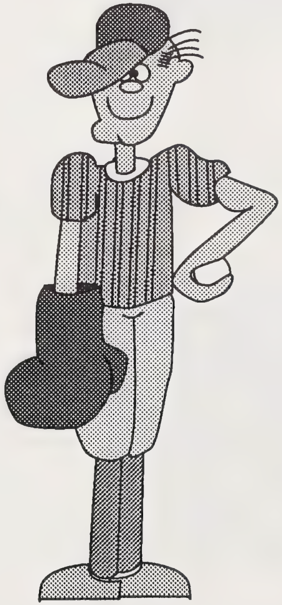
i. 6

ii. 11

iii. 22

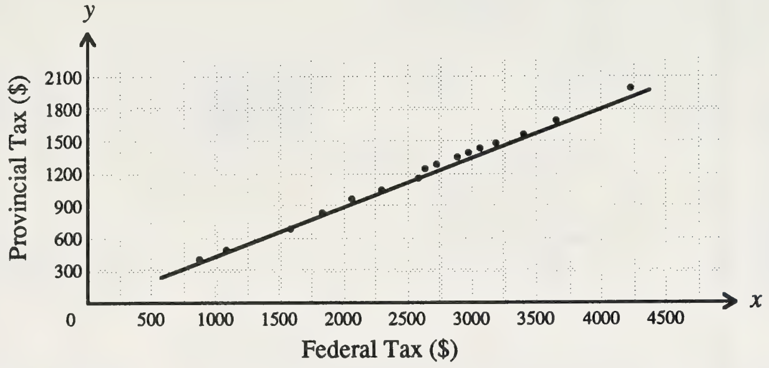
iv. 1

v. 30



3. Find the missing values using the following scatterplot.

Federal and Provincial Income Tax Payable in Northland
Before Surtax, 1989



a. What will be the federal tax when the provincial tax is the following?

i. $600

ii. $900

iii. $1800

iv. $2200

v. $3500

b. What will be the provincial tax when the federal tax is the following?

   i. $1000

   ii. $2500

   iii. $3700

   iv. $300

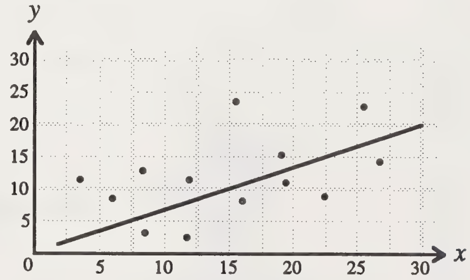
   v. $5500

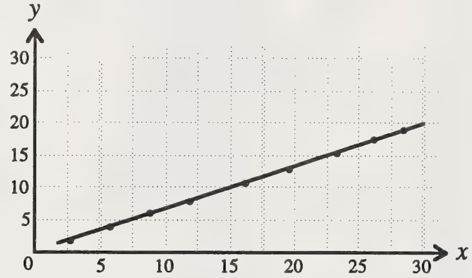
✓ For solutions to **Activity 3**, turn to **Appendix A, Topic 2.**

Activity 4

Determine the strength and type of correlation between the variables of a bivariate distribution.

Examine the lines of best fit for the following two scatterplots.





These two sets of data have the same line of best fit. However, the locations of the data are very different in the two scatterplots. The data in the second scatterplot is much closer to the line of best fit.

By examining the two scatterplots, you can also see that any data that is generated for the second scatterplot will be more reliable than the data generated for the first scatterplot. What you need to know is how much more reliable it will be.

What is needed is a way to tell how closely correlated the data is to the line of best fit. This can be done by calculating the **correlation coefficient**.

The **correlation coefficient** is a number that lies between 1 and – 1 inclusive.

If the coefficient is 1, there is a **perfect positive correlation** between the two sets of data. That is, the data will form a straight line that moves up to the right.

If the coefficient is – 1, there is a **perfect negative correlation** between the two sets of data. That is, the data will form a straight line that moves down to the right.

If the coefficient is 0, there is no correlation between the two sets of data.

The following formula can be used to calculate the correlation coefficient.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i is each value,

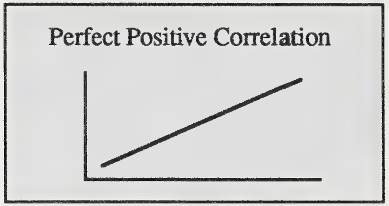
$\bar{x}$ is the mean,

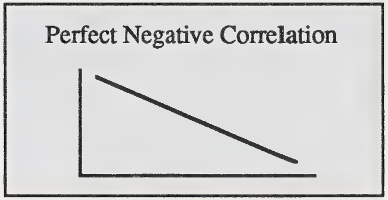
y_i is each value,

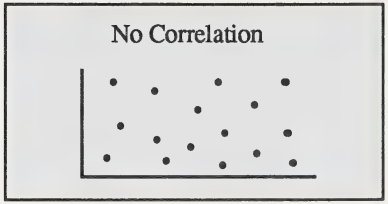
$\bar{y}$ is the mean

This looks like a complicated formula, but it becomes simple when you use the following steps to create a table. The Bases on Balls versus Strikeouts scatterplot will be used as an example.

A spreadsheet can be used to save you calculating time. See the **Extensions** section at the end of the topic.


Perfect Positive Correlation


Perfect Negative Correlation


No Correlation

Step 1: Write down the bivariate data in two columns.

Bases on Balls and Strikeouts by Pitchers in the Distance Baseball League, 1989

Pitcher	Bases on Balls	Strikeouts
Maddon, Calgary	81	140
Prune, Edmonton	44	131
Conic, Lethbridge	80	213
Daze, Medicine Hat	86	298
Garney, Grande Prairie	89	162
Drysday, High Level	50	127
Shoulder, Lloydminster	53	144
Jakson, Drumheller	71	161
Reussel, Red Deer	42	92
Hornblower, Jasper	73	178
Scottie, Banff	54	190
Suttuer, Calgary	70	144
Martten, Edmonton	55	120
Coodie, Lethbridge	57	175
Terence, Medicine Hat	34	65
Carrie, Grande Prairie	68	116
Smilie, Stettler	46	129
Mussen, St. Paul	58	112
Browner, St. Albert	64	124
Drowns, Wetaskiwin	47	118
Larry, Banff	56	180
Knocks, Red Deer	67	103

To save some space in writing down the information the number of bases on balls allowed will be labeled as x and the number of strikeouts will be labeled as y.

x	y
81	140
70	144
44	131
55	120
80	213
57	175
86	298
34	65
89	162
68	116
50	127
46	129
53	144
58	112
71	161
64	124
42	92
47	118
73	178
56	180
54	190
67	103

For your calculations it will not be important to distinguish between the bases on balls allowed and the strikeouts.

Step 2: Find the mean for each column. Round the means to the ones place and write them down in the next two columns.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$\bar{x} = \frac{81 + 70 + 44 + \ldots + 67}{22}$$

$$\bar{x} \doteq 61$$

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \ldots + y_n}{n}$$

$$\bar{y} = \frac{140 + 144 + 131 + \ldots + 103}{22}$$

$$\bar{y} \doteq 146$$

x	y	$\bar{x}$	$\bar{y}$
81	140	61	146
70	144	61	146
44	131	61	146
55	120	61	146
80	213	61	146
57	175	61	146
86	298	61	146
34	65	61	146
89	162	61	146
68	116	61	146
50	127	61	146
46	129	61	146
53	144	61	146
58	112	61	146
71	161	61	146
64	124	61	146
42	92	61	146
47	118	61	146
73	178	61	146
56	180	61	146
54	190	61	146
67	103	61	146

The lower the place value used for the mean, the more accurate the correlation will be. Here the mean is rounded to the ones place to keep the calculations simple.

Step 3: Take the difference of the values and the mean for each row. Write the solutions in the next two columns.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$
81	140	61	146	20	– 6
70	144	61	146	9	– 2
44	131	61	146	– 17	– 15
55	120	61	146	– 6	– 26
80	213	61	146	19	67
57	175	61	146	– 4	29
86	298	61	146	25	152
34	65	61	146	– 27	– 81
89	162	61	146	28	16
68	116	61	146	7	– 30
50	127	61	146	– 11	– 19
46	129	61	146	– 15	– 17
53	144	61	146	– 8	– 2
58	112	61	146	– 3	– 34
71	161	61	146	10	15
64	124	61	146	3	– 22
42	92	61	146	– 19	– 54
47	118	61	146	– 14	– 28
73	178	61	146	12	32
56	180	61	146	– 5	34
54	190	61	146	– 7	44
67	103	61	146	6	– 43

Step 4: Find the product of the rows in the last two columns. Write the product in the next column. Find the sum of this column.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
81	140	61	146	20	– 6	– 120
70	144	61	146	9	– 2	– 18
44	131	61	146	– 17	– 15	255
55	120	61	146	– 6	– 26	156
80	213	61	146	19	67	1273
57	175	61	146	– 4	29	– 116
86	298	61	146	25	152	3800
34	65	61	146	– 27	– 81	2187
89	162	61	146	28	16	448
68	116	61	146	7	– 30	– 210
50	127	61	146	– 11	– 19	209
46	129	61	146	– 15	– 17	255
53	144	61	146	– 8	– 2	16
58	112	61	146	– 3	– 34	102
71	161	61	146	10	15	150
64	124	61	146	3	– 22	– 66
42	92	61	146	– 19	– 54	1026
47	118	61	146	– 14	– 28	392
73	178	61	146	12	32	384
56	180	61	146	– 5	34	– 170
54	190	61	146	– 7	44	– 308
67	103	61	146	6	– 43	– 258
						9387

Step 5: Find the squares of each value in the two columns that contain the differences. Write these products in the next two columns. Find the sums of these two columns.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
81	140	61	146	20	– 6	– 120	400	36
70	144	61	146	9	– 2	– 18	81	4
44	131	61	146	– 17	– 15	255	289	225
55	120	61	146	– 6	– 26	156	36	676
80	213	61	146	19	67	1273	361	4 489
57	175	61	146	– 4	29	– 116	16	841
86	298	61	146	25	152	3800	625	23 104
34	65	61	146	– 27	– 81	2187	729	6 561
89	162	61	146	28	16	448	784	256
68	116	61	146	7	– 30	– 210	49	900
50	127	61	146	– 11	– 19	209	121	361
46	129	61	146	– 15	– 17	255	225	289
53	144	61	146	– 8	– 2	16	64	4
58	112	61	146	– 3	– 34	102	9	1 156
71	161	61	146	10	15	150	100	225
64	124	61	146	3	– 22	– 66	9	484
42	92	61	146	– 19	– 54	1026	361	2 916
47	118	61	146	– 14	– 28	392	196	784
73	178	61	146	12	32	384	144	1 024
56	180	61	146	– 5	34	– 170	25	1 156
54	190	61	146	– 7	44	– 308	49	1 936
67	103	61	146	6	– 43	– 258	36	1 849
						9387	4709	49 276

Step 6: You are now ready to substitute the numbers into the formula. The sum in the numerator of the formula is the sum of the seventh column and the two sums in the denominator of the formula are the sums of the eighth and ninth columns.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{9387}{\sqrt{[4709 \times 49\,276]}}$$

$$r_{xy} \doteq 0.62$$

This number indicates the strength and direction of the correlation of the two variates. Since the number is positive, the line of best fit is moving up to the right.

The number should only be used for comparison purposes. This correlation is stronger than a correlation of 0.3 but weaker than 0.75.

Answer any four of the following questions.

1. For each of the following, select the correlation that represents the scatterplot with the strongest correlation.

   a. – 0.64 and 0.64
   b. 0.3, 0.34, and 0.33
   c. 0.72, – 0.75, and 0.71
   d. 0.68, 0.86, and – 0.67

2. Find the correlation coefficient for the following bivariate data.

The Number of Wins and the Earned Run Average of Starting Pitchers in the Distance Baseball League, 1989

Pitcher	Wins	ERA
Maddon, Calgary	18	3.18
Prune, Edmonton	12	2.44
Conic, Lethbridge	20	2.22
Daze, Medicine Hat	13	3.67
Garney, Grande Prairie	12	3.69
Drysday, High Level	15	3.08
Shoulder, Lloydminster	16	3.26
Jakson, Drumheller	23	2.73
Reussel, Red Deer	19	3.12
Hornblower, Jasper	23	2.26
Scottie, Banff	14	2.92
Suttuer, Calgary	13	3.86
Martten, Edmonton	15	2.72
Coodie, Lethbridge	18	3.19
Terence, Medicine Hat	9	2.92
Carrie, Grande Prairie	10	4.29
Smilie, Stettler	13	3.25
Mussen, St. Paul	16	3.43
Browner, St. Albert	18	3.41
Drowns, Wetaskiwin	13	3.32
Larry, Banff	17	2.91
Knocks, Red Deer	14	3.14

3. Find the correlation coefficient for the following bivariate data.

The Mass and Fuel Consumption of Vehicles

Mass (kg)	Fuel Consumption (L/100 km)	Mass (kg)	Fuel Consumption (L/100 km)
800	5.0	1100	7.5
1300	7.8	1800	12.3
900	6.2	1400	10.4
1100	8.2	1700	12.0
1200	9.6	1000	8.5
600	3.9	1200	8.2
1500	9.8	900	5.6

4. Find the correlation coefficient for the following bivariate data.

Federal and Provincial Income Tax Payable for Northland
Before Surtax, 1989

Federal	Provincial	Federal	Provincial
820	381.20	1060	493.60
1540	716.40	1780	828.80
2040	949.90	2280	1060.30
2560	1190.10	2620	1218.20
2740	1274.40	2860	1330.60
2980	1386.80	3020	1404.20
3160	1469.10	3340	1553.40
3660	1702.60	4240	1972.90

5. Which set of bivariate data has the highest correlation?

For solutions to **Activity 4**, turn to **Appendix A, Topic 2**.

Activity 5

Collect, organize, and analyze sets of bivariate data.

Statistics can be viewed as a process that follows five basic steps.

- collection of data
- organization and presentation of data
- analysis of data
- drawing inferences
- evaluation for confidence

Look at how these steps are used in an actual situation.

Mr. Grant wanted to know how students' work habits influenced their grades.

Throughout the year, Mr. Grant gave the students two types of marks. The first mark showed the performance of the students on their work, assignments, projects, and tests. The second mark reflected the effort that the students put into their work. Mr. Grant thus completed the first of the five basic steps – collection of data.

Mr. Grant then organized the information into a chart. The chart had three columns: student identification, an average of the students' performance marks, and an average of the mark for student work habits. This information was then used to construct a scatterplot. In other words, Mr. Grant used the chart for organization and presentation of the data.

The line of best fit was found using the median method. Next, the equation of the line of best fit was found. This gave Mr. Grant a relationship between the performance and the work habits of the students. He could also use this equation to extrapolate new data. Thus, Mr. Grant's equation provided for an analysis of the data.

Mr. Grant then calculated the correlation coefficient. This gave Mr. Grant an indication of how well the line of best fit represented the data. Using this information, Mr. Grant was able to answer his original question. At this point, Mr. Grant would also analyze the situation and decide if one of the factors depends on the other factor and if there are any extraneous factors that influence both of these marks. At this point, therefore, Mr. Grant would be drawing inferences.

The evaluation for confidence will be discussed in the next topic.

You are now ready to do a project.

Perform the first four steps of statistics (collection of data, organization and presentation of data, analysis of data, and drawing inferences) on three of the following. Collect data from at least fifteen different people.

1. Compare the height of a person to the shoe size of that person.

2. Compare the height of a person to the waist measurement of that person.

3. Compare the height of a person to the finger stretch of that person. (Measure the finger stretch from finger tip to thumb tip across the back of the hand.)

4. Compare the height of a person to the hat size of that person. (For hat size, measure the circumference of the head.)

Which of the above has the highest correlation coefficient?

✓ For a typical solution to the first project, turn to **Activity 5, Appendix A, Topic 2.**

If you require help, do the Extra Help section.

If you want more challenging explorations, do the Extensions section.

} You may decide to do both.

Extra Help

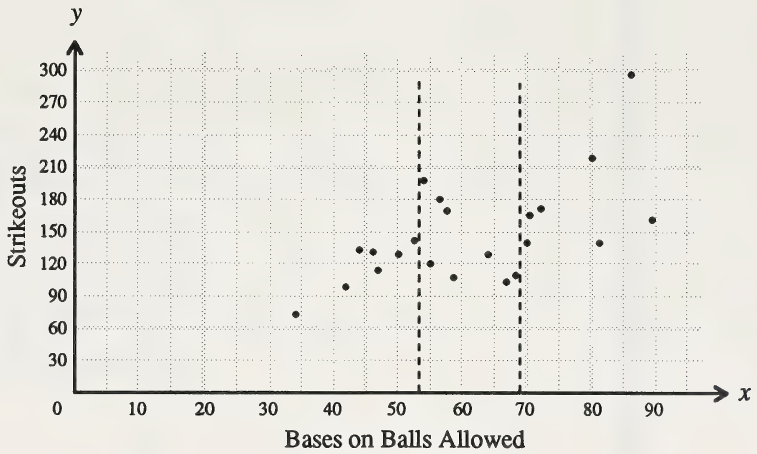
Finding the Line of Best Fit by the Inspection Method

Earlier in the topic, the median of each strip on the scatterplot was found by finding the median value of the *x*- and *y*-coordinates. This a long process. An alternate method is to find the median points by inspection.

The following are the steps that are used in finding the line of best fit using the inspection method.

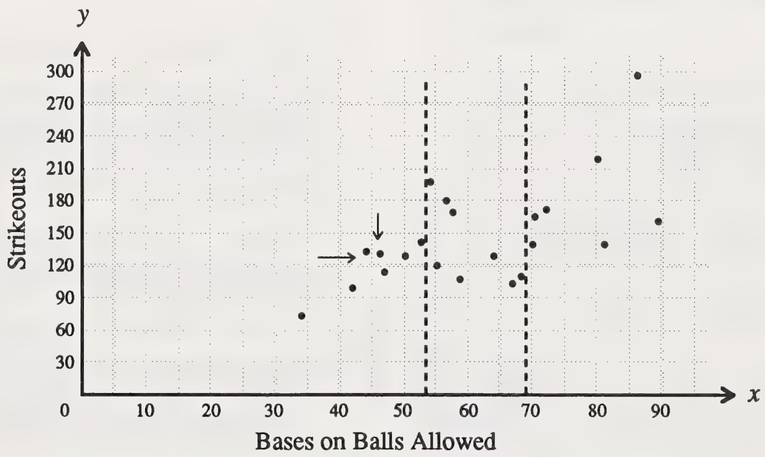
Step 1: Divide the scatterplot into three strips that have the same number of points in each strip. If it is not possible to put the same number of points in each strip, make sure that the first and last strips have the same number of points.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



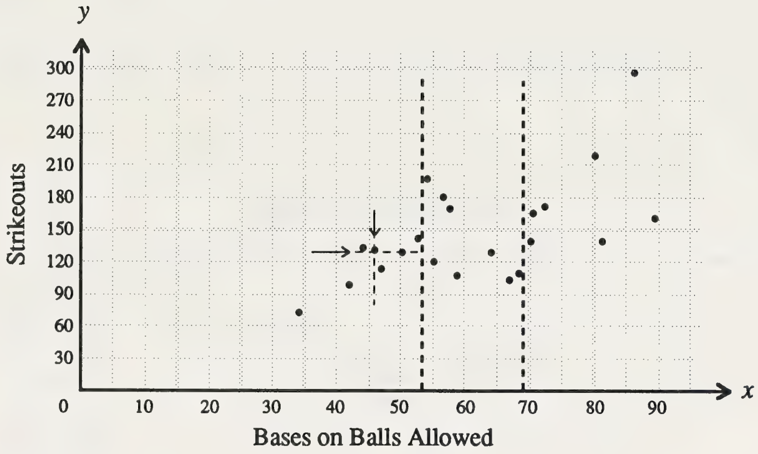
Step 2: Start with the first strip. If there are an odd number of points, select the middle point along the x-axis. If there are an even number of points, select a location halfway between the two middle points along the x-axis. Do the same for the y-axis.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



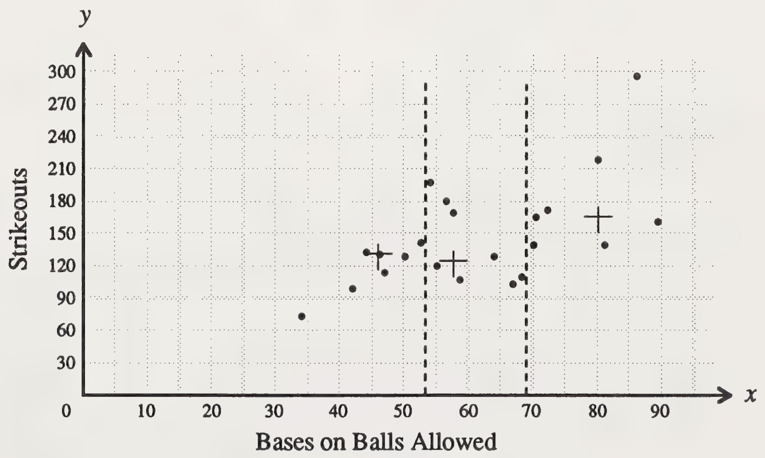
Step 3: Draw an imaginary line through the two arrows. The median point for the strip is located where the two lines intersect.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



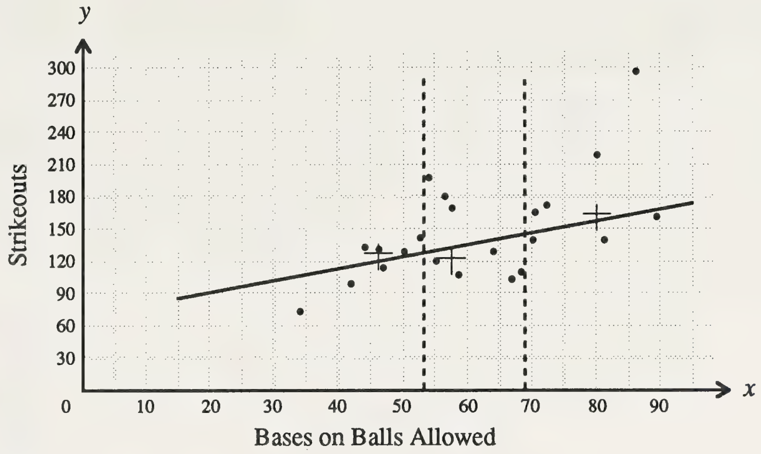
Step 4: Follow the same procedure as in Steps 2 and 3 and find the median points for the second and third strips.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989



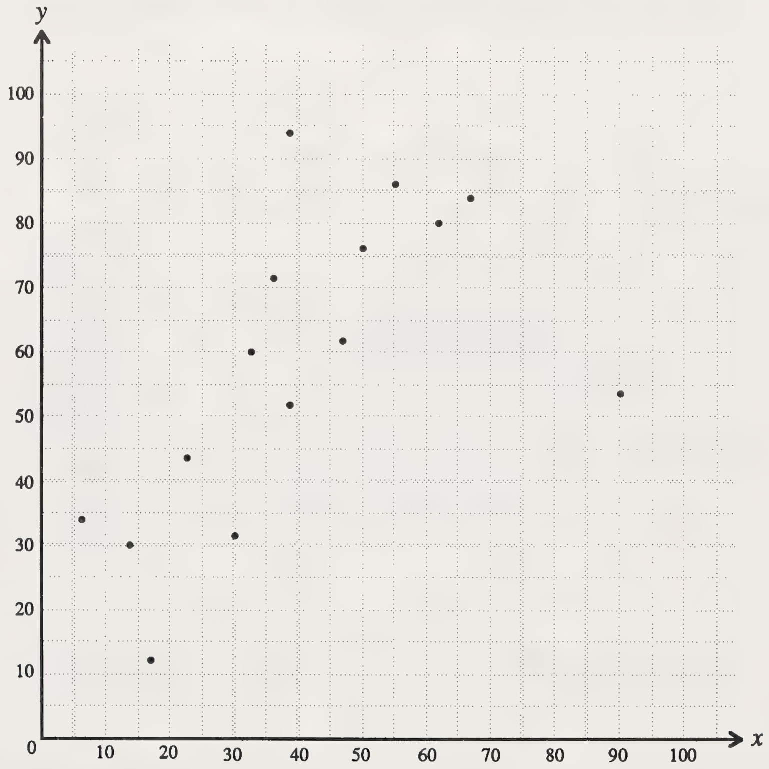
Step 5: The next step is to decide if the three median points lie on a straight line. You will notice if you use a ruler that these median points approximate a line. To determine the line, place your straightedge on the outside median points (median points in the first and last strips). Slide the straightedge a third of the distance to the middle median point (median point in the second strip) and draw the line. This will be the line of best fit.

Bases on Balls versus Strikeouts by Distance League Pitchers, 1989

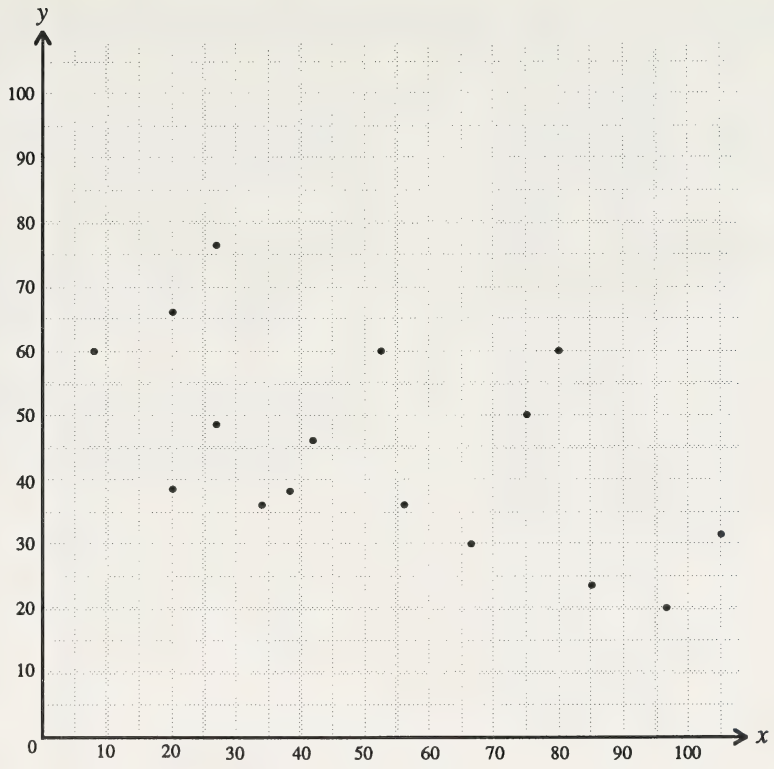


Try any two of the following questions.

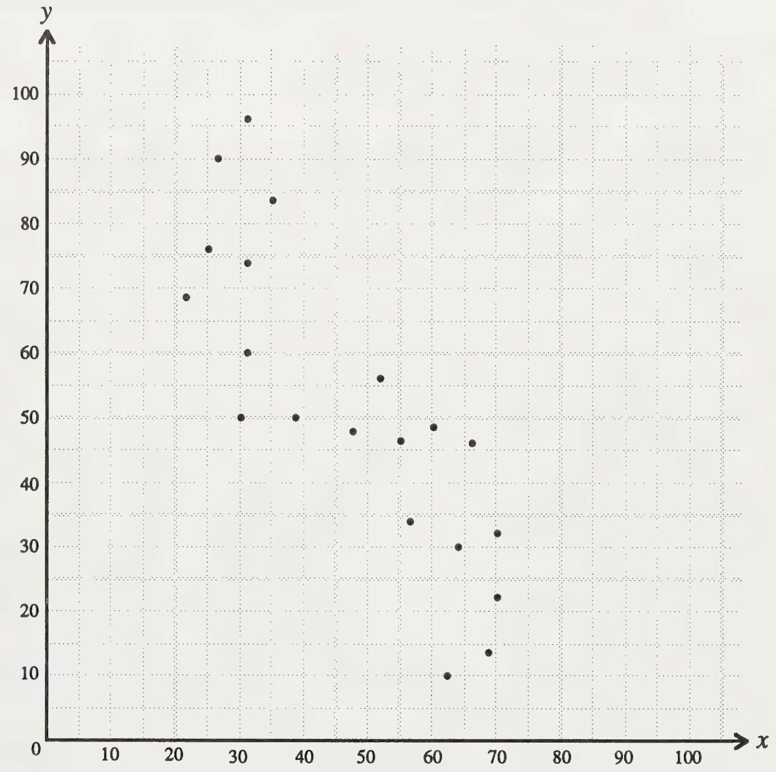
1. Use the inspection method to find the line of best fit for the following scatterplot.



2. Use the inspection method to find the line of best fit for the following scatterplot.



3. Use the inspection method to find the line of best fit for the following scatterplot.



For solutions to **Extra Help**, turn to **Appendix A, Topic 2**.

Extensions

Correlation Coefficient and the Spreadsheet

A lot of calculating time can be saved by using a spreadsheet. The spreadsheet will take the numbers and the equations that you give it and do all of the calculations for you. This really saves time.

The following data will be used for this spreadsheet. The *Appleworks* [1] spreadsheet for the Apple II [2] series computer will be used for this example.

This data looks similar to the previous data, but two values in the first column are different. Thus, the spreadsheet coefficient value may be slightly different than the coefficient obtained by the longhand method.

Bases on Balls and Strikeouts by Pitchers in the Distance Baseball League, 1989

Pitcher	Bases on Balls	Strikeouts
Maddon, Calgary	81	140
Prune, Edmonton	44	131
Conic, Lethbridge	80	213
Daze, Medicine Hat	86	298
Garney, Grande Prairie	89	162
Drysday, High Level	50	127
Shoulder, Lloydminster	53	144
Jakson, Drumheller	71	161
Reussel, Red Deer	42	92
Hornblower, Jasper	73	178
Scottie, Banff	53	190
Suttuer, Calgary	70	144
Martten, Edmonton	55	120
Coodie, Lethbridge	57	175
Terence, Medicine Hat	34	65
Carrie, Grande Prairie	70	116
Smilie, Stettler	46	129
Mussen, St. Paul	58	112
Browner, St. Albert	64	124
Drowns, Wetaskiwin	47	118
Larry, Banff	56	180
Knocks, Red Deer	67	103

[1] *AppleWorks* ™ is a trademark of Apple Computers, Inc.
[2] Apple II ™ is a trademark of Apple Computers, Inc.

Step 1: Write the bivariate data in the first two columns of the spreadsheet. Be sure to put the labels at the top of the columns. Since there are more pieces of data in the set than there are columns, you will not see all of the data at one time.

```
========A========B========C========D========E========
  1| x         y
  2|        81      140
  3|        70      144
  4|        44      131
  5|        55      120
  6|        80      213
  7|        57      175
  8|        86      298
  9|        34       65
 10|        89      162
 11|        70      116
 12|        50      127
 13|        46      129
 14|        53      144
 15|        58      112
 16|        71      161
 17|        64      124
 18|        42       92
-----------------------------------------------------------------------------------------
```

This is what you will see on the screen since the remaining lines are out of view.

Step 2: In cell C2 write down the formula @AVG(A2 . . . A23). This will calculate the mean of the values in the first column. Copy this formula to cell C3 through to cell C23.

In cell D2 write down the formula @AVG(B2 . . . B23). This will calculate the mean of the values in the second column. Copy this formula to cell D3 through to cell D23.

Be sure to label each of these columns as the mean.

```
========A========B========C========D========E========

 1| x           y          x-mean     y-mean
 2|          81       140  61.181818  146.45455
 3|          70       144  61.181818  146.45455
 4|          44       131  61.181818  146.45455
 5|          55       120  61.181818  146.45455
 6|          80       213  61.181818  146.45455
 7|          57       175  61.181818  146.45455
 8|          86       298  61.181818  146.45455
 9|          34        65  61.181818  146.45455
10|          89       162  61.181818  146.45455
11|          70       116  61.181818  146.45455
12|          50       127  61.181818  146.45455
13|          46       129  61.181818  146.45455
14|          53       144  61.181818  146.45455
15|          58       112  61.181818  146.45455
16|          71       161  61.181818  146.45455
17|          64       124  61.181818  146.45455
18|          42        92  61.181818  146.45455
----------------------------------------------------------------------------------------------
```

Note that the numbers go up to the twenty-third row since this is the last row that has a piece of the data.

This is what you will see on the screen since the remaining lines are out of view.

Since the computer can quickly handle calculations involving many decimal places, it is not necessary to round off the means.

Step 3: In cell E2 write the formula +A2 – C2. Copy this formula using the relative function through to cell E23. This column will find the difference between each of the values and the corresponding mean.

The same procedure will be performed in the next column to find the difference between the values and the mean for the second set of data. In cell F2 write the formula +B2 – D2. Copy this formula using the relative function through to cell F23.

```
=======A========B========C========D========E========F========
 1|  x           y           x-mean     y-mean     x – mean    y – mean
 2|         81        140  61.181818  146.45455   19.818182  –6.454545
 3|         70        144  61.181818  146.45455   8.8181818  –2.454545
 4|         44        131  61.181818  146.45455  –17.18182   –15.45455
 5|         55        120  61.181818  146.45455  –6.181818   –26.45455
 6|         80        213  61.181818  146.45455   18.818182   66.545455
 7|         57        175  61.181818  146.45455  –4.181818    28.545455
 8|         86        298  61.181818  146.45455   24.818182   151.54545
 9|         34         65  61.181818  146.45455  –27.18182   –81.45455
10|         89        162  61.181818  146.45455   27.818182   15.545455
11|         70        116  61.181818  146.45455   8.8181818  –30.45455
12|         50        127  61.181818  146.45455  –11.18182   –19.45455
13|         46        129  61.181818  146.45455  –15.18182   –17.45455
14|         53        144  61.181818  146.45455  –8.181818   –2.454545
15|         58        112  61.181818  146.45455  –3.181818   –34.45455
16|         71        161  61.181818  146.45455   9.8181818   14.545455
17|         64        124  61.181818  146.45455   2.8181818  –22.45455
18|         42         92  61.181818  146.45455  –19.18182   –54.45455
-----------------------------------------------------------------------------
```

This is what you will see on the screen since the remaining lines are out of view.

Step 4: In the next column place the products of the two differences. In cell G2 put the equation +E2*F2. Do a relative copy down the column.

```
========B========C========D========E========F========G========
  1|  y          x-mean     y-mean     x – mean   y – mean    Product
  2|       140  61.181818  146.45455  19.818182  –6.454545  –127.9173
  3|       144  61.181818  146.45455  8.8181818  –2.454545  –21.64463
  4|       131  61.181818  146.45455  –17.18182  –15.45455  265.53719
  5|       120  61.181818  146.45455  –6.181818  –26.45455  163.53719
  6|       213  61.181818  146.45455  18.818182  66.545455  1252.2645
  7|       175  61.181818  146.45455  –4.181818  28.545455  –119.3719
  8|       298  61.181818  146.45455  24.818182  151.54545  3761.0826
  9|        65  61.181818  146.45455  –27.18182  –81.45455  2214.0826
 10|       162  61.181818  146.45455  27.818182  15.545455  432.44628
 11|       116  61.181818  146.45455  8.8181818  –30.45455  –268.5537
 12|       127  61.181818  146.45455  –11.18182  –19.45455  217.53719
 13|       129  61.181818  146.45455  –15.18182  –17.45455  264.99174
 14|       144  61.181818  146.45455  –8.181818  –2.454545  20.082645
 15|       112  61.181818  146.45455  –3.181818  –34.45455  109.62810
 16|       161  61.181818  146.45455  9.8181818  14.545455  142.80992
 17|       124  61.181818  146.45455  2.8181818  –22.45455  –63.28099
 18|        92  61.181818  146.45455  –19.18182  –54.45455  1044.5372
----------------------------------------------------------------------
```

This is what you will see on the screen since the remaining lines are out of view.

Step 5: In the next two columns put the squares of the differences.

In cell H2 put the formula +E2^2 and in cell I2 put the formula +F2^2. Do a relative copy down both columns.

```
========D========E========F========G========H========I========
  1|  y-mean      x – mean    y – mean   Product     sq x – mean  sq y – mean
  2|  146.45455   19.818182   –6.454545  –127.9173   392.76033    41.661157
  3|  146.45455   8.8181818   –2.454545  –21.64463   77.760331    6.0247934
  4|  146.45455   –17.18182   –15.45455  265.53719   295.21488    238.84298
  5|  146.45455   –6.181818   –26.45455  163.53719   38.214876    699.84297
  6|  146.45455   18.818182   66.545455  1252.2645   354.12397    4428.2975
  7|  146.45455   –4.181818   28.545455  –119.3719   17.487603    814.84298
  8|  146.45455   24.818182   151.54545  3761.0826   615.94215    22966.025
  9|  146.45455   –27.18182   –81.45455  2214.0826   738.85124    6634.8430
 10|  146.45455   27.818182   15.545455  432.44628   773.85124    241.66116
 11|  146.45455   8.8181818   –30.45455  –268.5537   77.760331    927.47934
 12|  146.45455   –11.18182   –19.45455  217.53719   125.03306    378.47934
 13|  146.45455   –15.18182   –17.45455  264.99174   230.48760    304.66116
 14|  146.45455   –8.181818   –2.454545  20.082645   66.942149    6.0247934
 15|  146.45455   –3.181818   –34.45455  109.62810   10.123967    1187.1157
 16|  146.45455   9.8181818   14.545455  142.80992   96.396694    211.57025
 17|  146.45455   2.8181818   –22.45455  –63.28099   7.9421488    504.20661
 18|  146.45455   –19.18182   –54.45455  1044.5372   367.94215    2965.29752
-------------------------------------------------------------------------------
```

This is what you will see on the screen since the remaining lines are out of view.

Step 6: Find the sums of the last three columns.

In cell G24 put the formula @SUM(G2 . . . G23). Do a relative copy into cells H24 and I24.

```
========D========E========F========G========H========I========
  7|  146.45455   -4.181818  28.545455  -119.3719  17.487603  814.84298
  8|  146.45455   24.818182  151.54545  3761.0826  615.94215  22966.025
  9|  146.45455  -27.18182  -81.45455   2214.0826  738.85124  6634.8430
 10|  146.45455   27.818182  15.545455  432.44628  773.85124  241.66116
 11|  146.45455   8.8181818 -30.45455  -268.5537   77.760331  927.47934
 12|  146.45455  -11.18182  -19.45455   217.53719  125.03306  378.47934
 13|  146.45455  -15.18182  -17.45455   264.99174  230.48760  304.66116
 14|  146.45455  -8.181818  -2.454545   20.082646  66.942149  6.0247934
 15|  146.45455  -3.181818  -34.45455   109.62810  10.123967  1187.1157
 16|  146.45455   9.8181818  14.545455  142.80992  96.396694  211.57025
 17|  146.45455   2.8181818 -22.45455  -63.28099   7.9421488  504.20661
 18|  146.45455  -19.18182  -54.45455   1044.5372  367.94215  2965.2975
 19|  146.45455  -14.18182  -28.45455   403.53719  201.12397  809.66116
 20|  146.45455   11.818182  31.545455  372.80992  139.66942  995.11570
 21|  146.45455  -5.181818   33.545455 -173.8264   26.851240  1125.2975
 22|  146.45455  -8.181818   43.545455 -356.2810   66.942149  1896.2066
 23|  146.45455   5.8181818 -43.45455  -252.8264   33.851240  1888.2975
 24|                                    9281.1818  4755.2727  49271.455

-------------------------------------------------------------------------
```

The three sums found in this step are the three sums that are in the following correlation formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The sum that is found in cell G24 is the sum in the numerator of the formula. The sum found in cell H24 is the first sum in the denominator of the formula, and the sum found in cell I24 is the second sum in the denominator.

This leaves you with just one more step to find the correlation coefficient.

This is what you will see on the screen since the remaining lines are out of view.

Step 7: In cell I25 write the following formula: +(G24)/(H24*I24)^0.5.

This formula takes the sums that were found in the last step and puts them in the correlation coefficient formula. The number that appears in this cell is the correlation coefficient.

```
========D========E========F========G========H========I========
  8|  146.45455   24.818182  151.54545   3761.0826  615.94215  22966.025
  9|  146.45455  –27.18182  –81.45455   2214.0826  738.85124  6634.8430
 10|  146.45455   27.818182  15.545455   432.44628  773.85124  241.66116
 11|  146.45455   8.8181818 –30.45455  –268.5537   77.760331  927.47934
 12|  146.45455  –11.18182  –19.45455   217.53719  125.03306  378.47934
 13|  146.45455  –15.18182  –17.45455   264.99174  230.48760  304.66116
 14|  146.45455  –8.181818  –2.454545   20.082646  66.942149  6.0247934
 15|  146.45455  –3.181818  –34.45455   109.62810  10.123967  1187.1157
 16|  146.45455   9.8181818  14.545455   142.80992  96.396694  211.57025
 17|  146.45455   2.8181818 –22.45455  –63.28099   7.9421488  504.20661
 18|  146.45455  –19.18182  –54.45455   1044.5372  367.94215  2965.2975
 19|  146.45455  –14.18182  –28.45455   403.53719  201.12397  809.66116
 20|  146.45455   11.818182  31.545455   372.80992  139.66942  995.11570
 21|  146.45455  –5.181818   33.545455  –173.8264   26.851240  1125.2975
 22|  146.45455  –8.181818   43.545455  –356.2810   66.942149  1896.2066
 23|  146.45455   5.8181818 –43.45455  –252.8264   33.851240  1888.2975
 24|                                    9281.1818  4755.2727  49271.455
 25|                                    correlation coefficient 0.6063423

-----------------------------------------------------------------------
```

The correlation coefficient for this set of data rounds to 0.61.

This is what you will see on the screen since the remaining lines are out of view.

Do one of the following questions using the *AppleWorks* [1] spreadsheet.

1. Find the correlation coefficient for the following set of data.

Bridge Length versus Bridge Height

Bridge Length	Bridge Height
127 m	22 m
12 m	5 m
56 m	14 m
101 m	56 m
63 m	45 m
84 m	53 m
57 m	22 m
243 m	27 m
89 m	18 m
45 m	18 m
23 m	34 m
34 m	78 m
78 m	14 m
59 m	15 m
47 m	35 m
175 m	32 m
74 m	33 m
69 m	52 m

2. Find the correlation coefficient for the following set of data.

Goals Scored versus Assists Made by Minor League Hockey Players

Goals	Assists	Goals	Assists
82	53	74	57
52	63	50	65
49	57	47	51
46	48	42	49
39	48	37	42
36	41	36	38
35	37	34	37
34	35	27	38
25	39	22	39

✓ For solutions to **Extensions**, turn to **Appendix A, Topic 2**.

[1] *AppleWorks* ™ is a trademark of Apple Computers, Inc.

Topic 3 Confidence

Introduction

It is easy to draw conclusions and make inferences about different situations.

With the use of statistics you are now able to draw more reliable conclusions and make more accurate inferences about those situations.

This is still not enough. You need to know how accurate your conclusions and inferences are before the event happens.

The accuracy of your conclusions and inferences is referred to as the confidence of your conclusions and inferences and this is what you will be doing in this topic.

What Lies Ahead

Throughout this topic you will learn to

1. design and administer a yes/no simple survey and collect and organize the results of the survey
2. draw box and whisker plots of the results of multiple samples
3. use a 90% box and whisker plot chart to find the confidence interval for a survey result
4. draw statistical conclusions and make inferences to populations, and explain the confidence with which such conclusions and inferences are made based on the results of yes/no surveys
5. assess the strengths, weaknesses, and biases of given samples

Now that you know what to expect, turn the page to begin your study of confidence.

Exploring Topic 3

Activity 1

Design and administer a yes/no simple survey and collect and organize the results of the survey.

The Canadian government through Statistics Canada conducts two different types of census. The decennial census (conducted once every ten years) was first conducted in 1851 and is responsible by law to provide population counts of electoral districts to enable any redistributions of the seats in the House of Commons.

The quinquennial census (conducted once every five years) was first conducted in 1956 and is used to keep the nation's statistical information abreast of the demographic and socioeconomic developments that affect decision making in both the public and private sectors.

These two censuses are considered necessary to the well-being of the citizens of the nation.

Through the census, the decision makers of the nation can effectively plan to meet the needs of the citizens by the development and implementation of policies and programs.

Both of these censuses were most recently conducted in 1991.

The undertaking of these censuses is truly mammoth. Every household in the nation will receive census forms to be filled out and returned to Statistics Canada. Once the forms are returned the long task of compiling the information begins.

The information that has been collected by the censuses and other surveys conducted by Statistics Canada can be found in the Canada Year Book which has been published annually since 1959.

Often information is needed more quickly than it can be delivered by a census. At these times, a **survey** is taken. An example of a survey is the statistics compiled to determine the unemployment rate each month.

Gathering information has become so important to society that you will find other groups such as television and radio stations and newspapers that conduct surveys and opinion polls for their listeners and readers.

Since the information gathered by these censuses is so important to proper decision making in Canada, every household is required by law to complete and return the forms.

There are firms such as Gallup, Angus Reid, and Environics which make profits from conducting surveys for other groups.

This topic is not going to deal with the statistics that arise from the census, but instead the statistics that permit these firms to make statements about an entire group of people, or **population**, after they have only questioned a small group of these people, or a **sample** of the population.

Surveys or polls usually ask more than one question, and each question has the possibility of more than two responses. To help with your understanding of this material only single survey questions that have yes or no responses will be used. The following is one such question.

Do you approve of the way Brian Mulroney is handling the Canadian government?

There are only two ways to answer this question: yes or no.

This question was asked by an imaginary pollster, Runners. Look at the article that gave the results of this poll.

A survey in March of 1989 conducted by Runners asked 1257 adult Canadians the following question.

"Do you approve of the way Brian Mulroney is handling the Canadian government?"

Forty-three percent said that they approved. For results based on samples of this size, we are 95% confident that the error attributable to sampling and other random effects could be three percentage points above or below the given results.

The reader should also remember that the wording of the question and practical difficulties encountered in conducting the survey can introduce errors or bias into the findings of the survey.

This is a large survey and the method used is different from what you will learn in this topic.

What information is given in this article?

- The survey was conducted in March of 1989.
- The survey included 1257 adults.
- Forty-three percent of the sample said yes to the question.
- If the entire population was asked, 40% to 46% of the people would have responded yes 95% of the time.
- There may have been unknown circumstances that biased the findings.

A third response of undecided could also be used here. This response will be avoided to simplify the calculations.

It is fairly simple to see where the first three conclusions came from, but how did the survey firm arrive at the 95% confidence and the three percentage points above and below 43%?

Runners feels confident, based on its calculations, that if the entire population was surveyed, between 40% and 46% of the people would say yes to the question (3% above and below 43%). This statement is called a **statistical inference**, a prediction based on the gathered and organized data.

The 40% to 46% range (which is found by adding three to 43% and subtracting three from 43%) is called the **confidence interval**. The **confidence interval** is the range in which most of the samples would fall. (In this case, 95 out of every 100 samples would fall within the range.)

Now consider a smaller survey and assume the **population percentage** is 43%. The **population percentage** will be the percentage of the survey sample who responded yes.

Suppose you asked thirty people the survey question. How many would say yes to the question?

To perform this survey, you will have to make sure that the people are selected randomly. This is quite a task. You will have to make sure that people from all parts of the country are just as likely to be selected. You will have to make sure that all adult age groups are just as likely to be selected. You will have to make sure that men and women are just as likely to be selected. Also, other criteria will have to be considered such as religion, political affiliation, and socioeconomic status.

Instead of going to all of the trouble of trying to meet these criteria, a simulation will be used to generate the responses.

There are a number of ways that you can simulate this experiment. You could use a random number table, a spinner, the random number generator of a microcomputer, or you could draw marbles from a vat.

The 3% is called a sampling error. This sampling error is obtained from a formula, but this formula will not be discussed.

See the **Extensions** section at the end of this topic to see how to use a microcomputer to simulate these selections.

For this simulation, marbles will be drawn from a vat. The marbles will carry one of two labels, either yes or no.

To meet the conditions of the example, 43% of the marbles in the vat must have the label yes. Also, it will be necessary to make sure that the marbles are properly mixed and that they are the same size. Otherwise this will not be a random selection.

The population percentage can only be found by counting the number of yeses in the vat. If there are more than eight million marbles in the vat, it would be unreasonable to count all of the marbles. However, if a sample of the marbles is taken from the vat, the population percentage can be estimated by the **sample proportion**.

The **population percentage** will be the percentage of the objects in the vat that are **yeses** (labeled with the word yes).

The sample proportion can be found by dividing the number of yeses in the sample by the sample size.

Thirty marbles are removed from the vat. Of the thirty marbles, twelve are labeled with the word yes. Therefore, the sample proportion is $\frac{12}{30}$ or 0.4.

This sample proportion represents 40%. This lies within the 40% to 46% interval. If you returned the marbles to the vat, mixed them again, and then selected another thirty marbles, you would get another sample proportion.

Select thirty marbles from the vat forty times. Be sure that you replace the marbles and mix them in the vat after the results of each draw of thirty marbles are recorded.

The following is an example of these results.

Experiment	Responses	Number of Yeses
1	NYYNNNNNYNNYNNNYYNYYNNNNYNYNNN	10
2	YNYYNYNNYYNYYYYNNNNNYYYYYNNYYY	18
3	NNYNYYYNYYNYNYYNYYNNYNYYNYNYNN	16
4	NNYNNNYYYNYNYNNNNYNYYYYNNYNNNN	12
5	NYYNYYNNNNNYNNYYYYYYNNNNNNYNYY	14
6	NNNYNNNNNYYNNYYYNNNNNYYYNNNNNY	10
7	YYNYNNNNYNYNYYNYNNYNYNYYNNNNNY	13
8	YNNNYYNYYNNYYYYNNNYYYNNYNNYNNN	14
9	NNNNNNYNYYYYYNNNYNNNYYYYNNYNNY	13
10	NNYNYNYNYNNYYYYYNNNNNYYYNNYNNN	13
11	YNNYYYYYYNNYNYYNNNYNNNNYNNYNYN	14
12	YNNYNNNNNNYNNNNYNNYNNYYYNNNYNN	9
13	YNNYNNNNYNYYYNNYNNYYYNYYNYYNYN	15
14	YNYYNYNYYYNNYNYNNYNNYYNYYNNNNY	15
15	YNNYNNYYYNYNYNYNNYNNNNNNNYYNNY	12
16	YNNNNYYNNNNYNNNNYYYNNNNNNYNYNN	9
17	NNYNYNNNNNNNNNYNNYNNYYYYYYNYNN	11
18	YNNNNYYNNYNNYNNYYNYNNYNYNNYYNN	12
19	NYYNYYNYNNNYNYNNYNNNNYNYNNNNNY	11
20	YYYYYYNYNNNYNNNYYYNNNYYNYYNNNY	16
21	NNYNNYNNNYYNYYYNNYYNYYNNYYNYNY	15
22	NYYNNNYNYYYNYYNNYNYYYNYNNYYNNN	15
23	YYNNNYYYNYYYYYNYYNNYYNNNNNYYYY	18
24	YNNNNYNNNNYNYYYYNYYNNNNYYYNYNN	13
25	NNYYNYNNYYYYYNNNYYNNNNYNNNYYNN	13
26	NYNYNNNYYNNNYYNNYYNNYYYNYNNNNY	13
27	YYYNYNNNNNNYYNYYNYNNYYYNNYYNNN	14
28	NNNNYNYNNYNYNNNYNNYNYYNNYNNYYN	11
29	NNYNYYNYNYNNYNYNNYYYYYNYYYYYYN	18
30	YYNNNNNNNYNNNNNYYYNYNNNYNNNNNN	8
31	NNNNYYNYYNNNNNYYNNNNYYNYYNNYYN	12
32	YNNYNNNNYYYNYYYNNNYNYYYNYNNNYN	14
33	YYYYYYNYNNNYNYYYYYNYNNNNNYYNYY	18
34	NNNNYNYNNNNYYNNYNYNNNNNNNNYYYN	9
35	NNYNNNNYYYYNYNYNYNNYYNNNYNNYYY	14
36	NYNNNNNYNNYYNYYNYNNNNYNYYNNYYN	12
37	YNYYNNNYYNYNNYYNYYYYYYNNYYNNYY	18
38	YNNNYYNYNNNYNYYYNYYNNYYNNYYYYY	17
39	YNNYYNNYNNNYYNNNNYNYYYYNYNYNYY	15
40	NNYNNYNYNNNYYNYNNYNYNNNYNYYNYY	13

Once the data is collected, it should be sorted into a **sampling distribution**. The following sampling distribution has been collected from the table of forty groups of thirty marbles selected from the vat.

Number of Yeses	Sample Proportion	Tally	Frequency	Proportion of All Trials
0	0.00		0	0
1	0.03		0	0
2	0.07		0	0
3	0.10		0	0
4	0.13		0	0
5	0.17		0	0
6	0.20		0	0
7	0.23		0	0
8	0.27	|	1	0.025
9	0.30	|||	3	0.075
10	0.33	||	2	0.05
11	0.37	|||	3	0.075
12	0.40	|||||	5	0.125
13	0.43	||||| ||	7	0.175
14	0.47	||||| |	6	0.15
15	0.50	|||||	5	0.125
16	0.53	||	2	0.05
17	0.57	|	1	0.025
18	0.60	|||||	5	0.125
19	0.63		0	0
20	0.67		0	0
21	0.70		0	0
22	0.73		0	0
23	0.77		0	0
24	0.80		0	0
25	0.83		0	0
26	0.87		0	0
27	0.90		0	0
28	0.93		0	0
29	0.97		0	0
30	1.00		0	0
Total			40	1.00

Do you remember how the sampling distribution is constructed? If not, answer the following questions about the example results and the sampling distribution.

How many times were there ten yeses in the example results? What is the frequency for ten yeses from the sampling distribution? Both answers are 2.

How many times were there fifteen yeses in the example results? What is the frequency for fifteen yeses from the sampling distribution? Both answers are 5.

The frequency states the number of experiments that contain the specified number of yeses. For example, only one experiment contained seventeen yeses. Check the example results and sampling distribution to make sure that this is correct.

What is the most number of yeses that can be collected in each experiment? Since there are thirty responses for each experiment, the most number of yeses is 30.

What is 25 divided by 30 rounded to two decimal places? What is the sample proportion for twenty-five yeses from the sampling distribution? Both of these answers are 0.83.

What is 18 divided by 30? What is the sample proportion for eighteen yeses from the sampling distribution? Both of these answers are 0.60.

The sample proportion is the number of yeses divided by the total number of responses (or the highest possible number of yeses).

There was a total of forty experiments conducted for this example. The proportion of all trials is found by dividing the frequency by the total number of experiments, 40.

Examine this sampling distribution carefully. Does this distribution look familiar? The distribution resembles the frequency distribution table. This table keeps track of the number of samples that contained a specified number of yeses.

This sampling distribution is an approximation through simulation. The sampling distribution is the number of yeses in a sample of size 30.

This sampling distribution shows the amount of variability that exists between different random samples of a population. From this example, you would expect the number of yeses to be between 8 and 18 or 27% and 60%.

This confidence interval is much larger than the one given in the large survey of 1257 adults. Why?

This is not the exact sampling distribution for this situation. The sample should have a size of 1257. If you selected samples as large as those given in the example (1257), you are more likely to get the same confidence interval.

In general, to construct a sampling distribution, both the population percentage and the sample size must be known. With the sampling distribution, you can use the number of yeses in the sample or its equivalent form, the sample proportion, to determine the probability of getting a certain number of yeses or a certain sample proportion.

Do any three of the following questions.

1. Write survey questions to find the following information. The questions should only ask for yes and no responses.

   a. Do people prefer to watch hockey or baseball?

   b. Do people prefer large or small breakfasts?

   c. Do people prefer small or large cars?

   d. Do people prefer to vacation in Europe or in the sun on a Pacific island?

2. Write a series of yes/no survey questions that will find the following information.

   a. Which of the following fruits do people prefer to eat: bananas, oranges, or apples?

   b. For which of the following politicians would people vote to become prime minister: Brian Mulroney, Audrey McLaughlin, or Jean Chrétien?

3. Toss eight coins one hundred times. Record the results in a sampling distribution. Use your results to answer the following questions.

   a. What is the most likely number of heads each time you toss the eight coins?

   b. Estimate the probability of getting 2, 3, 7, and 8 heads? (Give four answers.)

   c. Estimate the probability of getting 2, 3, 7, or 8 heads? (Give one answer.)

4. Explain how you would set up a simulation for the following situation: Out of a sample of twenty people, fourteen people prefer to vacation in the Caribbean.

✓ For solutions to **Activity 1**, turn to **Appendix A, Topic 3**.

Activity 2

Draw box and whisker plots of the results of multiple samples.

In the last activity you constructed a sampling distribution. In this activity you are going to take that sampling distribution and use the information to construct a 90% box and whisker plot.

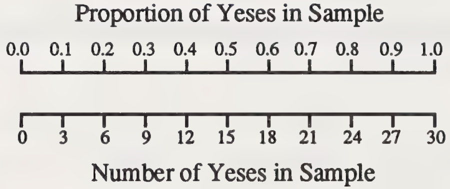
The sampling distribution that you constructed in Activity 1 came from the following survey question:

> Do you approve of the way Brian Mulroney is handling the Canadian government?

This same sampling distribution is shown.

The following steps are used to construct a box and whisker plot for the sampling distribution.

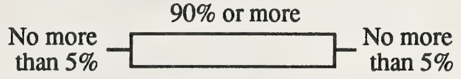
Step 1: Construct two axes: the first one is for the proportion of yeses in the sample, and the second one is for the number of yeses in the sample.



Number of Yeses	Sample Proportion	Tally	Frequency	Proportion of All Trials
0	0.00		0	0
1	0.03		0	0
2	0.07		0	0
3	0.10		0	0
4	0.13		0	0
5	0.17		0	0
6	0.20		0	0
7	0.23		0	0
8	0.27	|	1	0.025
9	0.30	|||	3	0.075
10	0.33	||	2	0.05
11	0.37	|||	3	0.075
12	0.40	|||||	5	0.125
13	0.43	||||| ||	7	0.175
14	0.47	||||| |	6	0.15
15	0.50	|||||	5	0.125
16	0.53	||	2	0.05
17	0.57	|	1	0.025
18	0.60	|||||	5	0.125
19	0.63		0	0
20	0.67		0	0
21	0.70		0	0
22	0.73		0	0
23	0.77		0	0
24	0.80		0	0
25	0.83		0	0
26	0.87		0	0
27	0.90		0	0
28	0.93		0	0
29	0.97		0	0
30	1.00		0	0
Total			40	1.00

Most surveys use a 95% box and whisker plot. The 90% box and whisker plot is done here since it is easier to calculate.

Step 2: Find the number of values that will be in the box and the whiskers. Since you are making a 90% box and whisker plot, 90% of the values must be in the box and 5% of the values will be in each whisker. The box can have more than 90% of the values but any one whisker will have at most 5% of the values.



In this particular case, there are forty trials. There will be
90% of $40 = 0.9 \times 40 = 36$
values in the box. Also, there will be
5% of $40 = 0.05 \times 40 = 2$
values in each of the whiskers.

Step 3: Find the number of yeses for these locations. The completed diagram for this box and whisker plot is in Step 5, so refer to it as you read these instructions for Step 3.

The location of the left endpoint for the left whisker is 8 or 0.27. See the diagram in Step 5.

The location of the point where the left whisker joins the box is 9 or 0.3. The left end of the box starts at 9 or 0.3 since the frequency for nine yeses is 3 and the left whisker can have no more than two values. The left whisker already has 1 for the frequency which represents the eight yeses.

The location of the point where the right whisker joins the right end of the box is 18 or 0.6. The reason is that the frequency is 5 for the eighteen yeses and these five cannot be separated. Thus, the box contains thirty-nine values. This means that there is no right whisker on the diagram. The location of the right endpoint for the right whisker can be interpreted as being at 18 or 0.6.

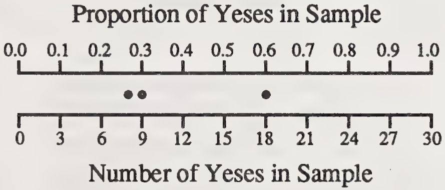
As you can see from these results, the endpoint for the right whisker and the end of the right side of the box will be the same point.

For a 95% box and whisker plot, you put 95% of the values in the box and 2.5% of the values into each whisker.

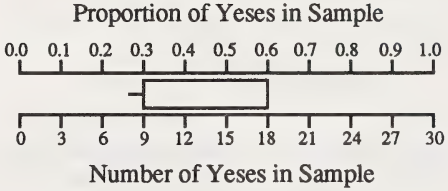
Because of these ties in the sample proportions, you cannot get exactly thirty-six values in the box. This box actually contains thirty-nine values, counting the edges of the box. When a tie occurs, such as here, you will always construct the plot so that no more than 5% of the values are in either whisker. The box can have more than 90% of the values, but any one whisker will have at most 5% of the values.

The case you are doing now is a good example of this. There will be only one point outside of the box on the left side, which is 2.5% of the values. Also, there will be no values outside the box on the right side, or 0% of the values. This means that there is 97.5% of the values in the box.

Step 4: Find the ends of the box and the whiskers on the grids. Mark these locations off with dots.



Step 5: Using the dots as your guides, draw in the box and the whiskers.



The box and whisker plot is now complete.

If you put the left edge of the box at 10 instead of 9, you would have four values in the whisker. This is unacceptable since you are allowed only two values in the left whisker.

If you would like more work on creating 90% box and whisker plots, go to the **Extra Help** section at the end of the topic.

Do either one of the following questions.

1. Make a 90% box and whisker plot for the following sampling distribution.

Number of Yeses	Sample Proportion	Tally	Frequency	Proportion of All Trials
0	0.00		0	0
1	0.05		0	0
2	0.10		0	0
3	0.15		0	0
4	0.20		0	0
5	0.25		0	0
6	0.30		0	0
7	0.35		0	0
8	0.40		0	0
9	0.45		1	0.017
10	0.50		5	0.083
11	0.55		11	0.183
12	0.60		13	0.217
13	0.65		12	0.2
14	0.70		10	0.17
15	0.75		6	0.1
16	0.80		2	0.03
17	0.85		0	0
18	0.90		0	0
19	0.95		0	0
20	1.00		0	0
Total			60	1.00

2. Make a 90% box and whisker plot for the following sampling distribution.

Number of Yeses	Sample Proportion	Tally	Frequency	Proportion of All Trials
0	0.00		0	0
1	0.05		0	0
2	0.10		0	0
3	0.15		0	0
4	0.20		0	0
5	0.25		0	0
6	0.30		0	0
7	0.35		0	0
8	0.40		0	0
9	0.45		1	0.01
10	0.50		4	0.04
11	0.55		9	0.09
12	0.60		13	0.13
13	0.65		18	0.18
14	0.70		23	0.23
15	0.75		17	0.17
16	0.80		8	0.08
17	0.85		4	0.04
18	0.90		3	0.03
19	0.95		0	0
20	1.00		0	0
Total			100	1.00

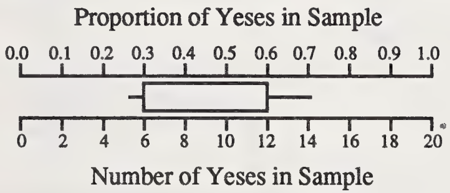
✓ For solutions to **Activity 2**, turn to **Appendix A, Topic 3**.

Activity 3

Use a 90% box and whisker plot chart to find the confidence interval for a survey result.

In the last activity you constructed a box and whisker plot that contained 90% (or more) of the pieces of samples in the box. In this activity you are going to learn how to read charts of 90% box and whisker plots.

Start by taking a closer look at a box and whisker plot that contains 90% of the values in the box.

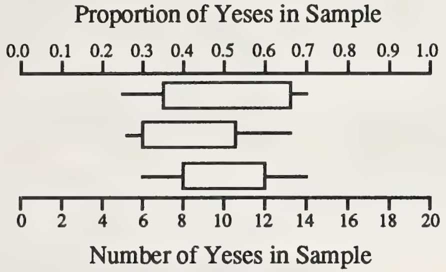


This 90% box and whisker plot is constructed to contain the middle 90% of the sample proportions in the box. The sample proportions that fall within the box will be called **likely sample proportions.** (The edges of the box will be considered as being inside the box.) The sample proportions that fall within the whiskers will be called **unlikely sample proportions.**

If you were to take further samples, it is possible that you could get a sample proportion that will fall outside of the whiskers. Sample proportions that fall outside of the whiskers will also be called **unlikely sample proportions.** Therefore, unlikely sample proportions are sample proportions that fall outside of the box.

In the box and whisker plot shown, the likely sample proportions will be **between** 0.3 and 0.6 inclusive. The unlikely sample proportions will be **less than** 0.3 and **more than** 0.6.

This box and whisker plot was constructed using a population percentage of 40% and forty samples of size 20. The following box and whisker plots were also constructed using the same dimensions.



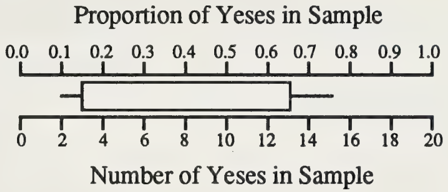
What do these cases tell you?

Each time that you perform the experiment, you get different data; therefore, you get different box and whisker plots.

How can you get the box and whisker plots to be the same?

The key here is in the number of samples. In these cases, there were forty samples of size 20 selected. As the number of samples is increased for each of these experiments, the more these box and whisker plots will become similar.

The following is the box and whisker plot for the same experiment, but this time 1000 samples were used.

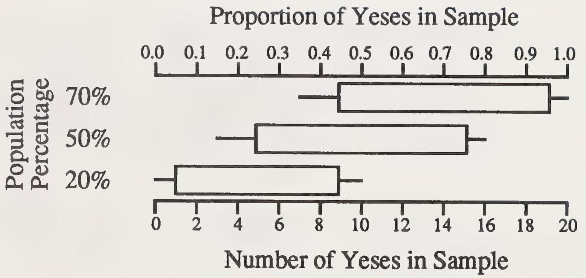


As each of the previous experiments **increases** its number of samples (but keeps the same sample size), the closer its box and whisker plot will come to this box and whisker plot. The **more samples** of a particular size, the **more accurate** the box and whisker plot will be.

What effect does the population percentage have on the box and whisker plot?

Forty samples of size 20 can be explained as follows. Forty samples means the experiment was repeated forty times, while size 20 refers to the number of selections. An example might be selecting twenty marbles from a container that contains a large number of marbles having only two different colours. The twenty marbles are replaced in the container after the results of each draw of twenty are recorded.

Examine the following box and whisker plots.

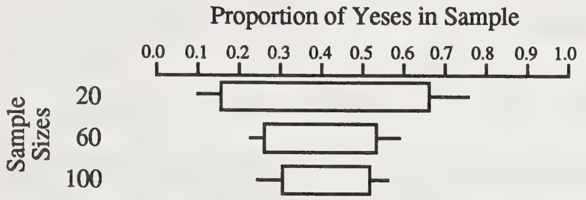


All three of these box and whisker plots are using 1000 samples of size 20.

As the population percentage **increases**, the box and whisker plot **moves** to the right. Therefore, for every different population percentage, there is a different box and whisker plot.

What effect does the size of the sample have on the box and whisker plot?

Examine the following box and whisker plots that have a population percentage of 40% and are using 1000 samples of different sizes.



The number of yeses in the sample are not given here since each sample size has a different number of yeses for each proportion.

As the sample size gets **larger**, the box becomes **shorter**. Therefore, for every different sample size, there is a different box and whisker plot.

In summary, every different population percentage and every different sample size will require a different box and whisker plot.

A number of these box and whisker plots have already been constructed for you and appear in **Appendix B**. Take a look at them. Notice that there are five different charts. Each chart represents one of the following sample sizes: 20, 40, 60, 80, or 100. Within each chart there are twenty-one different population percentages represented from 0% to 100%.

Take some time to see how you can use these charts.

About 35% of the mathematicians in Canuckland are women. If a random sample of twenty mathematicians is taken, will the following results be likely or unlikely?

- twelve men and eight women
- four men and sixteen women
- eighteen men and two women

Step 1: Identify the chart that you will be using.

The five charts in **Appendix B** are distinguished by the sample size. In this particular example, a sample of size 20 is being selected since a sample of twenty mathematicians is taken. It is shown on the next page.

Step 2: Select the appropriate parts of the chart.

You are given two different types of information in the question.

- a population percentage
- three different numbers that represent the number of yeses in the sample

First, find 35% on the left or right side of the chart. Move horizontally across the chart to the 90% box and whisker plot that represents a population percentage of 35%. This is the only box and whisker plot that you will be using for this question.

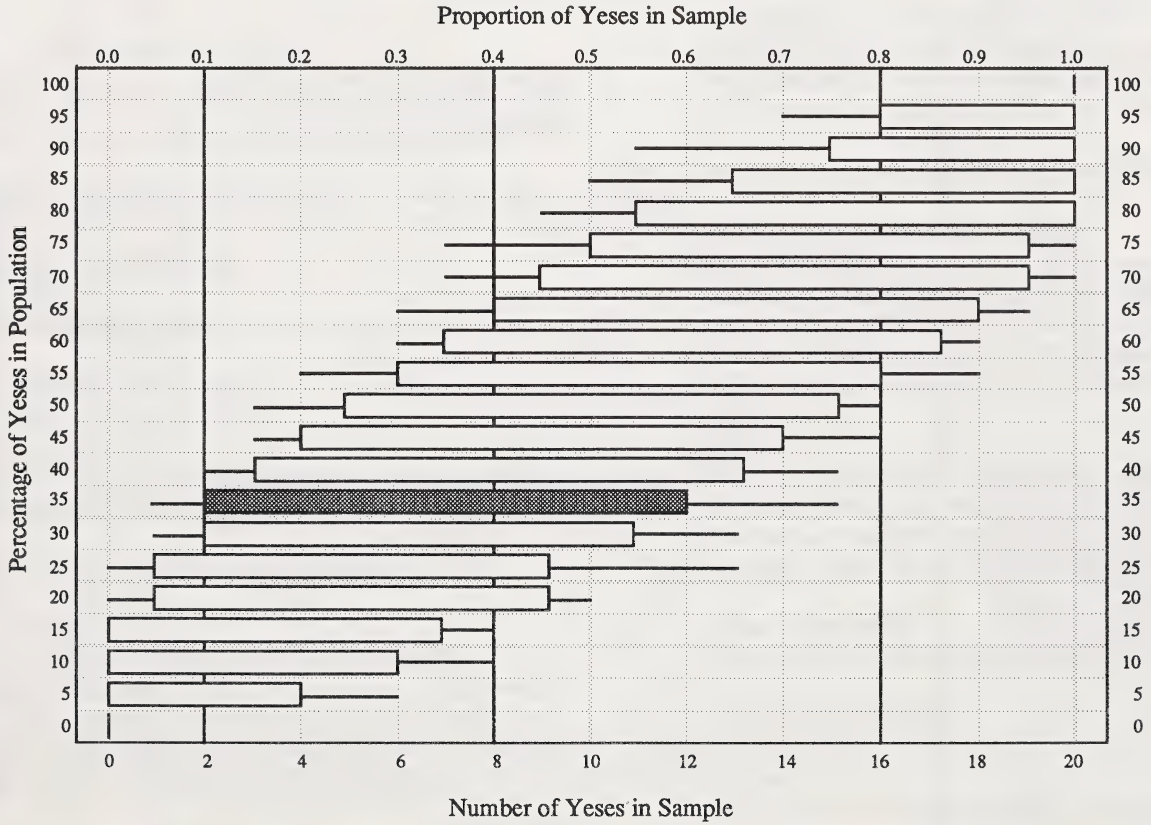
Next, find the three different numbers on the chart that represent the number of yeses in the sample (the number of women in the three questions). These three numbers will be along the bottom of the chart. Since this question deals with women, ignore the numbers that deal with men and use only the numbers that apply to women.

Step 3: Decide if the different samples are likely or unlikely.

Move straight up from each of these three points. If the path of this movement goes through the box of the 35% box and whisker plot, the sample is likely. If the path does not go through the box, the sample is unlikely.

This information is summarized on the following chart. Notice that the 35% box and whisker plot has been shaded. Lines have been used to identify the paths from the number of yeses. These paths represent the number of women for each question.

90% Box and Whisker Plots from Samples of Size 20



The first path up from eight yeses goes through the box. Therefore, a sample with eight women is likely.

The second path up from sixteen yeses does not go through the box. Therefore, a sample with sixteen women is unlikely.

The third path up from two yeses goes through the left edge of the box. Since the edges of the box are considered as part of the box, a sample with two women is likely.

Take a look at a different type of situation.

A pharmaceutical company has data from a sample of twenty people in which fourteen of the people responded favourably. What population percentages are likely for this situation?

Step 1: Identify the chart that you will be using.

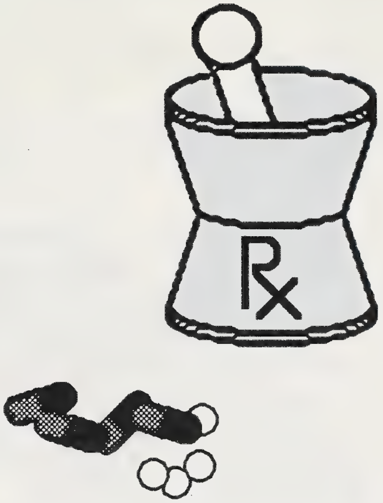
The five charts in **Appendix B** are distinguished by the sample size. In this particular example, a sample of size 20 is being selected.

Step 2: Select the appropriate parts of the chart.

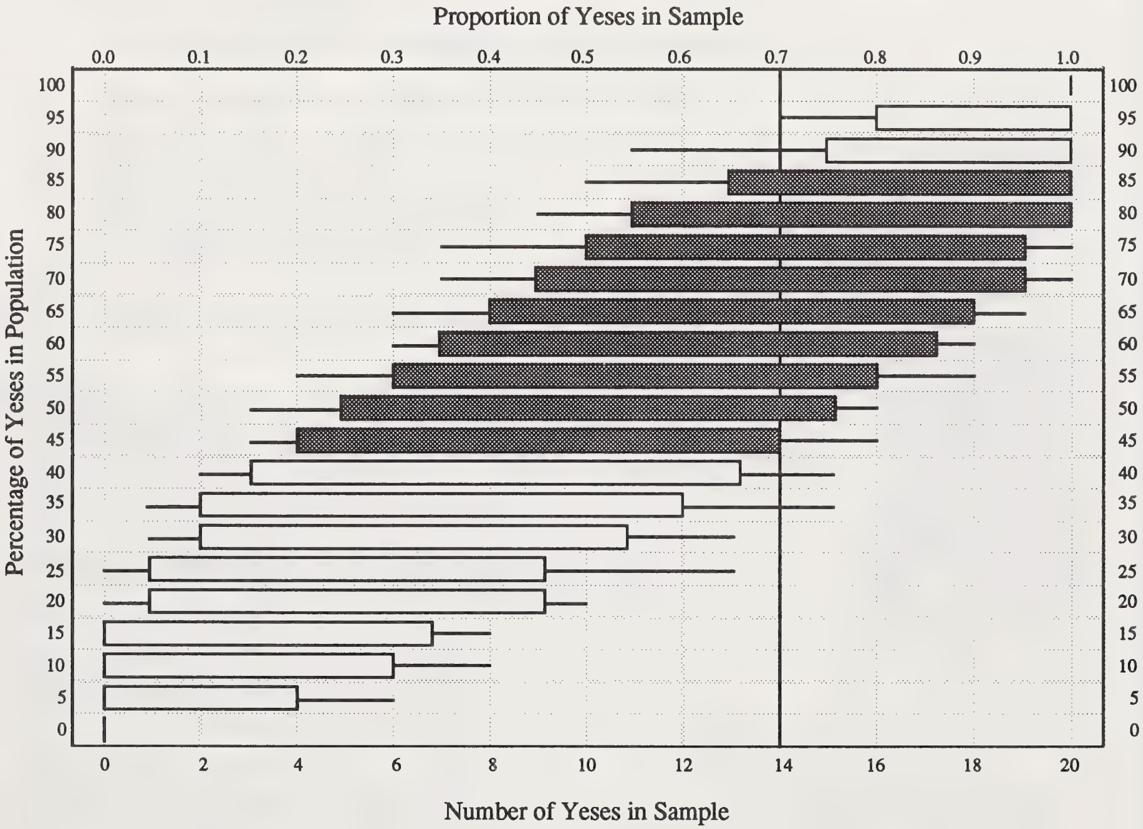
In this particular question, the only part that needs to be identified is the number of yeses, which is fourteen.

Step 3: Move straight up and identify the population percentages that are likely for this number of yeses.

Move straight up from the fourteen yeses. All of the boxes that the path passes through will be likely population percentages.



90% Box and Whisker Plots from Samples of Size 20



The likely population percentages are 45% to 85% inclusive.

This question has demonstrated the central idea of this topic. The range of population percentages is called the **confidence interval**.

The 90% confidence interval for the pharmaceutical company is 45% to 85%.

What does this mean? If the pharmaceutical company surveyed a sample of twenty people, it would be 90% certain that 45% to 85% of the people would say yes.

The procedure for finding the confidence interval is exactly the procedure that was performed in the last example.

The following are the steps that you will use for finding the confidence interval.

- Identify the appropriate number of yeses or sample proportion.
- Draw a straight vertical line from this point.
- Any box that the line passes through will be part of the confidence interval. Remember that the edges of the boxes are part of the box.
- In any concluding statement, make sure to state that out of every 100 trials, 90 trials will succeed and 10 trials will fail.

Try a third and different situation.

What are the likely sample proportions if a sample size of 60 is drawn from a population of 65% yeses?

Step 1: Identify the chart that you will be using.

The five charts in **Appendix B** are distinguished by the sample size. In this particular example, a sample of size 60 is being selected.

Step 2: Select the appropriate parts of the chart.

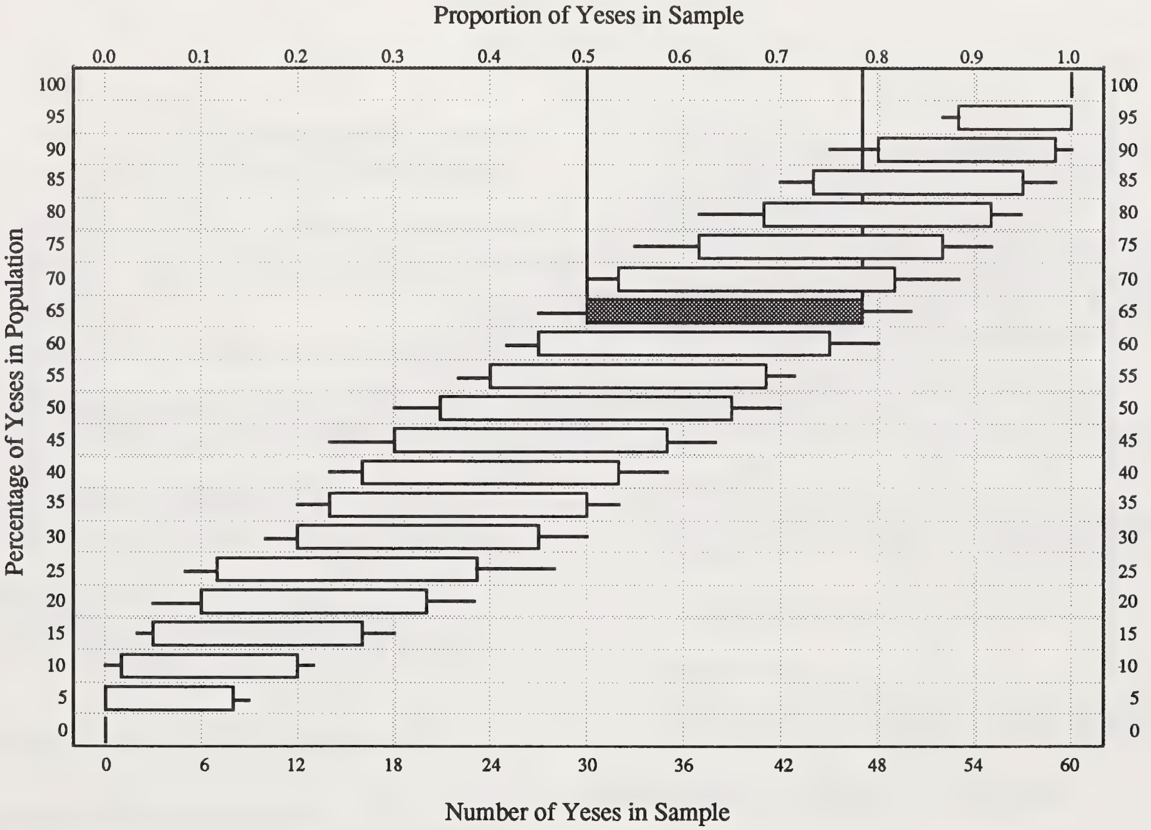
In this particular question, the only part that needs to be identified is the population percentage box and whisker plot for 65%.

Step 3: Move straight up from both the left and right of the box and identify the appropriate sample proportions.

This movement is shown in the following chart. The 65% box and whisker plot has been highlighted. Since you are only interested in the sample proportions, it is not necessary to move down to the number of yeses scale. Therefore, the lines only move up to the top scale.

The sample proportions identified in the chart are from 0.5 to 0.78.

90% Box and Whisker Plots from Samples of Size 60



Do any seven of the following questions.

1. A sample size of 40 is selected from a population of 90% yeses. Are each of the following sample proportions likely or unlikely?

   a. 0.25 yeses
   b. 0.65 yeses
   c. 0.825 yeses
   d. 1.00 yeses

2. A random sample of size 80 contains a sample proportion of 0.35 yeses. For which of the following population percentages is this a likely sample proportion?

   a. one with 15% yeses
   b. one with 20% yeses
   c. one with 25% yeses
   d. one with 30% yeses
   e. one with 35% yeses
   f. one with 40% yeses
   g. one with 45% yeses
   h. one with 50% yeses
   i. one with 55% yeses
   j. one with 60% yeses

3. A random sample of size 60 contains seventeen yeses. For which population percentages is this a likely sample proportion?

4. What are the likely sample proportions if you draw a random sample of size 100 from a population with 75% yeses?

5. A recent survey showed that 65% of all motor vehicle accidents on the Great Plains involved alcohol. If a random sample of sixty drivers involved in accidents were taken, determine the following:

   a. Is it likely that a sample proportion of 0.45 will have alcohol in their blood?

   b. Is it likely that forty-eight drivers will have alcohol in their blood?

6. Seventy percent of the people who live in Centreland believe that there should be a national sales tax. If you take a random sample of twenty people from Centreland, determine the following:

   a. Is it likely or unlikely that a sample proportion of 0.95 will approve of the tax?

   b. Is it likely or unlikely that ten people will approve of the tax?

   c. Is it likely or unlikely that a sample proportion of 0.35 will approve of the tax?

7. In a survey, 45% of the people said they would buy a certain new car model constructed by the Canuckland Engineering Corporation. If a sample size of 40 is used, what percentage of the population will buy the car?

8. What are the confidence intervals for the following situations?

   a. a sample size of 20 and a sample proportion of 0.55

   b. a sample size of 100 and 83 yeses

   c. 70 yeses for a sample size of 80

✓ For solutions to **Activity 3**, turn to **Appendix A, Topic 3**.

Activity 4

Draw statistical conclusions and make inferences to populations, and explain the confidence with which such conclusions and inferences are made based on the results of yes/no surveys.

In this activity you are going to employ the tools that you have learned to use in the last three activities to draw conclusions and inferences about real-world situations.

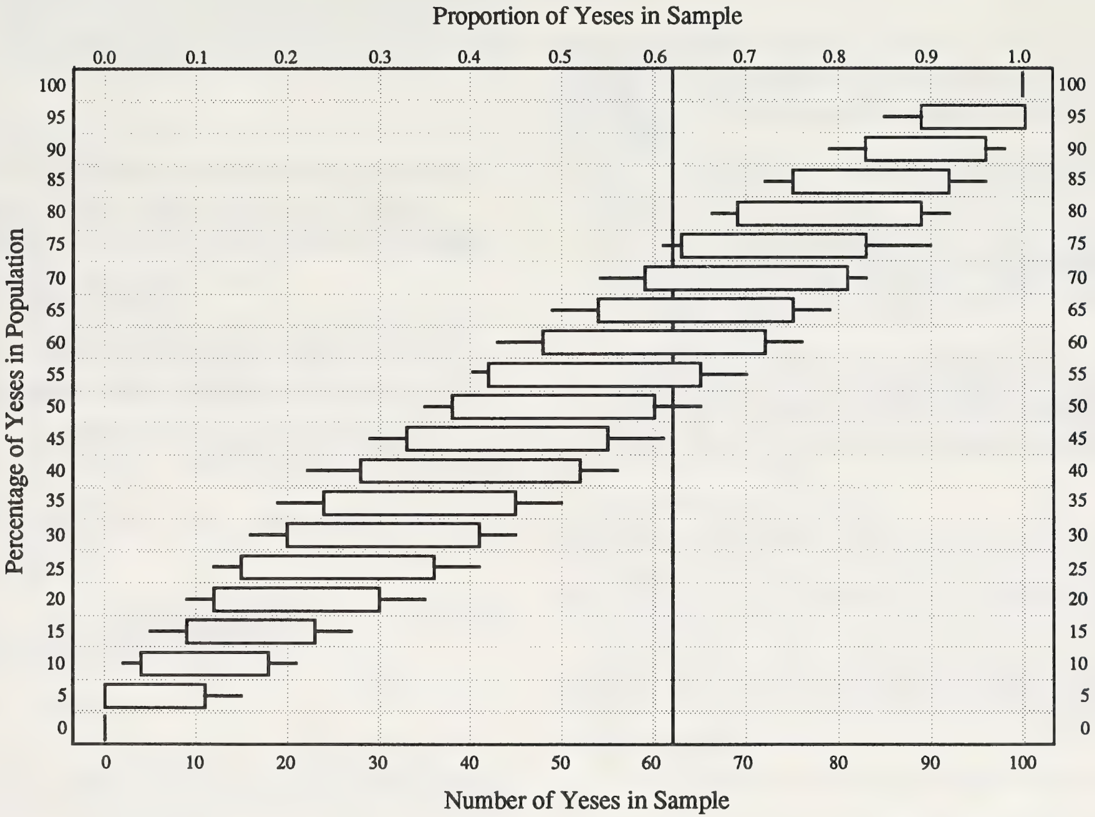
Go through an example and see how these tools are used to draw a proper conclusion.

Student Power magazine gave 100 students two different erasable pens to use for a week: Tic Trick and Matey Peel. The students used the two pens for a week while doing their usual schoolwork. At the end of the week, Student Power surveyed the students to discover which pen they preferred. According to the results, sixty-two students preferred the Tic Trick over the Matey Peel pen.

Assuming that the students were selected randomly, what conclusions would you draw?

First, examine the box and whisker chart for a sample size of 100. Drawing a line through 62 yeses, you find the confidence interval. For this particular case, the line goes through the boxes for the population percentages 55% to 70%. This chart is as follows.

90% Box and Whisker Plots from Samples of Size 100



Then the following conclusion is drawn.

> If all of the 100 students of a random sample were asked whether they preferred the Tic Trick pen or the Matey Peel pen, 55% to 70% of the students would select the Tic Trick 90% of the time. If 100 groups of 100 students were asked, 90 groups would have 55% to 70% of the students choosing the Tic Trick over the Matey Peel.

The concluding statement to this situation and all concluding statements must have the following information.

- the size of the sample
- the confidence interval
- how many times this confidence interval will happen

In this topic you have been working with 90% confidence intervals. This means that your results will be wrong ten times out of every 100 trials. For the 95% confidence intervals, the results will be wrong only five times out of every 100 trials.

The concluding statements that you are most likely to see will deal with 95% confidence intervals. These confidence intervals are calculated using the same procedure that you used in calculating 90% confidence intervals.

Try any five of the following questions.

1. If you take a random sample of size 20 from a population with 35% yeses, is 0.70 a likely sample proportion?

2. In order to be 90% confident, how many samples out of 500 will be likely?

3. If you want to increase the length of the confidence interval, should you increase or decrease the sample size?

4. Fast Food Inc. surveyed a group of sixty people. The company asked the people if they ate breakfast every morning. Write the concluding statement for the survey if forty-nine of the people responded yes to the question.

5. Runners surveyed a group of forty people. The people were asked if they favoured running the public schools year-round. Write the concluding statement for the survey if seventeen of the people responded yes to the question.

6. A group of eighty people was surveyed to determine whether they wanted more parks. Write the concluding statement for the survey if seventy-three of the people responded yes to the question.

For solutions to **Activity 4**, turn to **Appendix A, Topic 3**.

Activity 5

Assess the strengths, weaknesses, and biases of given samples.

In this activity you are going to be looking at the different ways errors can be made in surveys.

Most errors are made during the sampling process. You will examine different types of sampling and the procedures that should be followed when you are conducting a survey.

Sampling

Throughout this topic it has been important that the samples were collected randomly. If a **random sample** was not selected, the method of constructing confidence intervals would not be legitimate.

What is a **random sample**? For a sample to be selected randomly, the following two conditions must be met.

- Each member of the population must have the same chance or probability of being selected.
- Each member of the sample must be selected independently of the others.

It is important to note that these conditions control how the sample is selected, not the results of the selection.

Look at an example.

Twenty students are going to be selected from a school of 300 students.

The name of each student is placed on a 5 cm by 10 cm card. The cards are placed in a large box and shaken. A blindfolded person selects twenty cards from the box, one card at a time.

Is this a random selection? This question can be answered by checking to see if the two conditions have been met.

- Since all of the cards were the same size and the selector had no way of distinguishing among the cards, every student had the same chance of being selected.
- Since only one card was selected each time, there was no way the selection of one card influenced the selection of another card. Therefore, each selection was made independently of the other selections.

Therefore, this is a random selection.

Now suppose that all of the students selected were members of the swimming club. Were the students still selected randomly?

Of course they were. The conditions of random selection were met as shown. The results of the selection do not determine whether a random selection was made.

Since the students are being selected randomly, you would expect the sample to be representative of the population. That is, if all of the twenty students were members of the swimming club, then you would expect all of the students in the population to be members of the swimming club.

Making a random selection is not as easy as it sounds. Suppose you want to take a random sample of the citizens of the city of Red Deer. To make sure that you get a random sample, you will need the names of everyone who lives in Red Deer.

Compiling this material will be very time-consuming. You will have to go to every household in the city and find the names of everyone who lives there. There will be times when no one is home. There will be times when people will give you incorrect information or no information. To make this even worse, people will be moving between places in the city, which may cause you to miss some people or put them on the list twice. In addition, people may be moving in and out of the city, causing some people to be left off the list and other people to be on the list who no longer live in the city.

All of these problems will prevent you from getting a true random sample.

Any list that you find for the people in the city will also have these same problems and others.

> Voting lists will list only those people who were eligible to vote at the time of the last election.
>
> City household lists will list only the households; they will not state how many people are in each family or their names.
>
> Telephone directories will list only those individuals who have phones and want their names listed in the directory.

When a method of selection tends to overrepresent or underrepresent some part of the population, the method is said to **bias** the sample. As shown, it is almost impossible to make a selection without some bias. If it is impossible for you to make a selection without a bias, then make the selection as randomly as possible.

Since it is not possible to get a truly random sample of the entire population, different types of sampling procedures are used to gather a sample. Now you will look at some of these sampling techniques and see how they can be used to create a biased sample.

Convenience Sampling: Each member of the sample is simply chosen without the use of any random mechanism.

A convenience sample could be selected by picking the first twenty people you see, or the first twenty people who talk to you, or the twenty tallest people, etc.

Since convenience sampling uses no explicit random mechanism to select the sample, no confidence interval statements can be made about the population.

Convenience samples are usually biased.

Self-Selected Sampling: Each member of the sample has volunteered to be in the survey. Self-selected sampling is a type of convenience sampling.

The most common type of self-selected sampling is returning a survey form from a newspaper or magazine.

People who volunteer to participate in a survey usually are the most concerned and motivated people in the population.

Judgement Sampling: Each member of the sample has been selected by an expert or someone who is knowledgeable about the population. Judgement sampling is a type of convenience sampling.

Judgement sampling is done by a produce purchaser. The purchaser will select and taste several fruits from a shipment in order to decide on the quality of the shipment. This type of sampling sometimes can give you a better measurement of the quality of a population than a random sample.

Probability Sampling: Each member of the sample is selected by use of a probability mechanism. Also, each member of the population may not have the same chance of being selected or of being selected independently.

Probability sampling may or may not be a random sample. This decision is made with the selection of the probability mechanism.

Statisticians have developed methods for calculating confidence intervals from probability samples, but these methods are complicated and will not be discussed in this course.

Remember: Your work with the confidence interval depended on the sample being selected randomly.

Clustered Sampling: A group is selected randomly from the population and then every member of the group becomes a member of the sample. This is a type of probability sampling, but it is not simple random sampling.

A town may use a clustered sample by randomly selecting streets, then randomly selecting households along those streets, and finally surveying everyone who lives in those households.

Clustered sampling is not a type of convenience sampling since at no time does the surveyor decide who will be included in the sample. Since clusters or groups are being selected, not all of the members of the population have the same chance of being selected. Also, since the group is bound together by some common trait, the members of the sample are not being selected independently.

Stratified Random Sampling: Each member of the population is placed into a strata or subgroup, and then the members of the sample are found by taking a random selection from each strata or subgroup. This is a type of probability sampling, but it is not simple random sampling.

An example of stratified random sampling occurs when a school is divided into a group of boys and a group of girls, and then twenty people are selected from each of the two groups. Another example occurs when the school is divided into its grades, and then twenty students are selected from each grade.

A stratified sample may be used when you want to know how a certain strata replies in comparison to the whole population. As an example, you may want to know not only how the whole student body answers a survey, but how both the boys and the girls answer that survey.

A stratified sample also may be used when you want to ensure that certain strata have their opinions heard. This will cause the sample to be more representative of the entire population. The increased representation causes the confidence interval for the stratified sample to be shorter than the confidence interval from a random sample of the same population. This will cause the stratified sample to give more precise estimates than a random sample.

Systematic Sampling: Each member of the sample is selected from the population by using a system. This is a type of probability sampling.

One type of systematic sampling is to number each member of the population, and then select each *n*th (fourth, eighth, or tenth, for example) member of the population to be part of the sample. An example would be to select every fifth student who walks into your classroom.

Systematic sampling has several advantages over random sampling. It is easier to do than random sampling and it guarantees that the sample is taken from throughout the population. There are disadvantages to systematic sampling. It is possible that certain characteristics may be able to hide in the population using this method. An example of this would be the case where a street intersection is being studied. If the intersection is only studied on Monday and Tuesday, the study will not cover the traffic problems that show up on Friday and Saturday.

Systematic sampling is most useful when random sampling is too difficult to be carried out and the ordering of the population does not create a problem.

Other Forms of Bias

Even if an organization uses random sampling or probability sampling to choose a sample, the survey can become biased for other reasons.

- People may refuse or neglect to respond.
  - The question might be sensitive to some people.
  - The interviewer may not be able to contact some people.
- People may be untruthful.
  - People may be trying to give socially acceptable responses.
  - People might be trying to tell the interviewer the answers that they believe the interviewer wants to hear.
  - People might be trying to hide the fact that they do not know the answer to the question.
  - People may inadvertently improve or change a numerical response.
- The survey may be conducted at a poor time.
  - There may have just been a campaign concerning the topic of the survey.
- The survey may be poorly worded.
  - People are more willing to accept certain phrases. For example, people are more willing to accept the phrase "not allow" than "forbid."
  - The surveyor may word the question to cause people to respond in a certain way.
- The interviewers may not be properly trained.
  - An interviewer can easily misinterpret people's responses.
- Errors may be made in repetitive tasks such as computer entry errors or simple tabulation errors.

The following are some hints for making your survey as unbiased as possible.

- Make sure that a good cross section of the population is selected for the survey.
- Do a follow-up to the survey to make sure that all the people in the sample have responded.
- Take extreme care in the wording of the questions to eliminate all ambiguities and to avoid using socially sensitive words.
- Ensure anonymity in the survey. Some people will not respond if their responses can be identified.
- Make sure that the question asks for a yes/no response. Many people will not take the time to write out responses.
- In interviews, the interviewer must be careful not to emphasize any one word or phrase. If the respondent fails to understand the question, the interviewer must only repeat the question and not explain or interpret the question.

Answer any five of questions 1 to 6, and question 7.

1. Which of the following situations is a random sampling of students from a high school? Explain why any situation is not a random sampling.

   a. Select the students whose last digit in their phone number is a 4.

   b. Assign each student a number and then use a random number table to select the students.

   c. Select only those students whose name starts with the letters A, B, C, or D.

2. A sample of 200 people is selected from across Canada. Three people live on the same street. Is it possible that this is a random sample? Explain your answer.

3. Explain why you would not be able to obtain a true random sample of the people of Calgary.

4. State what type of sampling is being used in each of the following cases.

   a. Readers are asked to return a survey that was given in a magazine.

   b. One student from each home classroom in a school is selected to be on a social committee.

   c. A buyer for a cannery tries several different pieces of fruit to judge its quality.

5. Explain three different situations in which a person might give an untruthful answer to a survey.

6. Explain why a survey might report that 5% more women than men responded that they were married.

7. Plan and carry out your own survey.

   a. Write a yes/no-type survey question on a topic that you find interesting.

   b. Decide what your population will be. As an example, you may use your school or your community as your population.

   c. Try your survey question on a few people that you know to make sure that they are interpreting the question as you planned.

   d. Decide on a method and select a random sample of forty people. Explain your method.

   e. Survey the forty people in your sample. Tabulate the results.

   f. What is the sample proportion for your survey?

   g. What is the 90% confidence interval for the sample proportion of your survey?

   h. Write a concluding statement to explain the findings of your survey.

✓ For solutions to **Activity 5**, turn to **Appendix A, Topic 3**.

If you require help, do the Extra Help section.

If you want more challenging explorations, do the Extensions section.

} You may decide to do both.

Extra Help

Constructing a 90% Box and Whisker Plot Chart

In Activity 2, you were given the following steps for creating a 90% box and whisker plot.

Step 1: Construct two axes. The first one is for the proportion of yeses in the sample, and the second one is for the number of yeses in the sample.

Step 2: Find the number of values that will be in the box and the whiskers. Since you are making a 90% box and whisker plot, 90% of the values will be in the box and 5% of the values will be in each whisker on both sides of the box. The box can have more than 90% of the values, but any one whisker will have at most 5% of the values.

Step 3: Identify the number of yeses that will mark the end of the box and the whiskers.

Step 4: Find the ends of the box and the whiskers on the axes. Mark these locations off with dots.

Step 5: Using the dots as your guides, draw in the box and the whiskers.

The following is part of the sampling distribution for a population percentage of 5% yeses. The five steps shown will be used to make the 90% box and whisker plot for this information.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	355	0.355
1	0.05	320	0.32
2	0.10	179	0.179
3	0.15	92	0.092
4	0.20	52	0.052
5	0.25	1	0.001
6	0.30	1	0.001
7	0.35	0	0
8	0.40	0	0
9	0.45	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

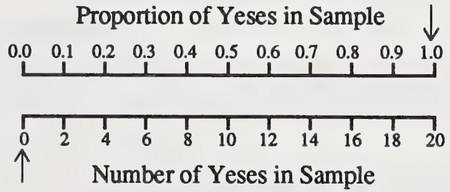
Step 1

Construct two axes. The first one is for the proportion of yeses in the sample, and the second one is for the number of yeses in the sample.

The axis for the proportion of yeses will represent what percentage of the sample is made up of yeses, where 0 will represent the case where there are no yeses in the sample, and 1 will represent the cases where the entire sample is made of yeses.

The second axis will represent the number of yeses in the sample.

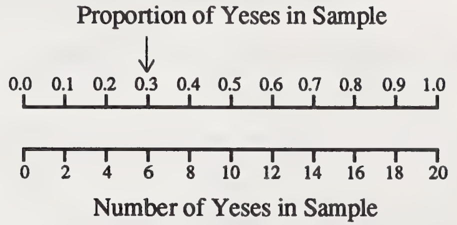
Since there are twenty responses in the sample, twenty yeses will correspond to the number 1 on the proportion of yeses scale. Therefore, you must make sure that these two numbers line up on the chart.



Also note that the zeros on both grids line up.

Choose any number along the bottom scale. Divide it by 20. Find this number along the top scale. Go straight down to the bottom scale. Is this the number that you chose? It should be.

For example, start with the number 6. Dividing 6 by 20 will give you 0.3. Find 0.3 along the top axis. From here move straight down to the bottom axis. Are you back at the number 6? Check the following axes.



The axes are now ready.

Step 2

Find the number of values that will be in the box and the whiskers. Since you are making a 90% box and whisker plot, 90% of the values will be in the box and 5% of the values will be in each whisker on both sides of the box.

The sampling distribution has 1000 values. If 90% of the values are to be in the box, then 90% of 1000 or 900 values will be in the box.

That leaves 100 values outside of the box, or 50 values on each side of the box.

Step 3

Identify the number of yeses that will mark the ends of the box and the whiskers.

Count down from the top of the frequency column in the sampling distribution until you get to the number 50. The rows you have covered in the sampling distribution will represent the left whisker of the plot. If the row that you are in adds up to 51 or more, that category will be part of the box. Remember that the box must contain at least the middle 900 of the values from the sampling distribution table. The box can contain more than 900 of the values, but it must not contain fewer.

Since the first row of this box contains more than fifty values in the first row, this plot will not have a left whisker. The box will start at a point on the far left of the axis.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	355	0.355	Since this row already contains more than fifty values, there will be no left whisker.
1	0.05	320	0.32	
2	0.10	179	0.179	
3	0.15	92	0.092	
4	0.20	52	0.052	
5	0.25	1	0.001	
6	0.30	1	0.001	

Now count up from the bottom of the frequency column. Again you want the frequencies to add up to 50, but not more than 50. These rows will represent the right whisker of the plot.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	355	0.355	
1	0.05	320	0.32	
2	0.10	179	0.179	
3	0.15	92	0.092	
4	0.20	52	0.052	
5	0.25	1	0.001	These two rows will represent the right whisker. The next row will not be part of the whisker since it will add up to 54, which is more than 50.
6	0.30	1	0.001	

The remaining rows will be part of the box.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	355	0.355	These rows will be represented by the box in the chart.
1	0.05	320	0.32	
2	0.10	179	0.179	
3	0.15	92	0.092	
4	0.20	52	0.052	
5	0.25	1	0.001	
6	0.30	1	0.001	

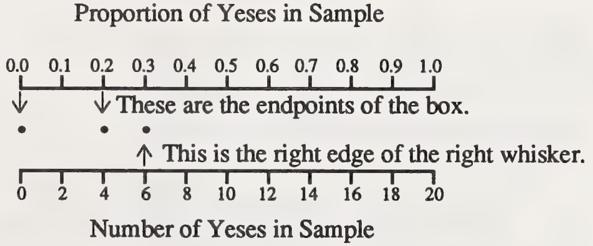
From this you can see that the left edge of the box will be at 0 yeses, the right side of the box will be at 4 yeses, and the right edge of the right whisker will be at 6 yeses.

Now you are ready to move on to the fourth step.

Step 4

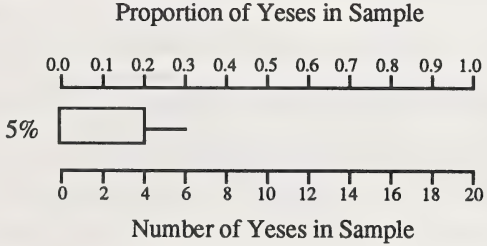
Find the ends of the box and the whiskers on the axes. Mark these locations off with dots.

This particular grid will have only three dots. The first two dots will mark the locations of the edges of the boxes and the third dot will mark the endpoint of the right whisker.


Proportion of Yeses in Sample
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
These are the endpoints of the box.
This is the right edge of the right whisker.
0 2 4 6 8 10 12 14 16 18 20
Number of Yeses in Sample

Step 5

Using the dots as your guides, draw in the box and the whiskers.



The 90% box and whisker plot for the 5% population percentage is now complete.

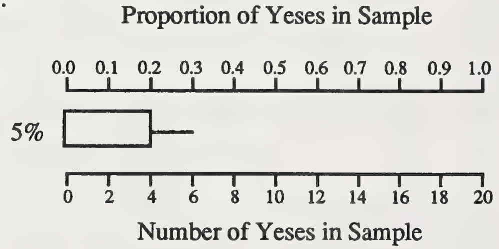
The following sampling distribution is used to find the 90% box and whisker plot for a population percentage of 10%. Use this information and draw the 90% box and whisker plot.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	126	0.126
1	0.05	214	0.214
2	0.10	219	0.219
3	0.15	180	0.180
4	0.20	121	0.121
5	0.25	87	0.087
6	0.30	39	0.039
7	0.35	12	0.012
8	0.40	2	0.002
9	0.45	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

Step 1

Construct two axes. The first one is for the proportion of yeses in the sample, and the second one is for the number of yeses in the sample.

You will graph the 10% yeses plot with the 5% yeses plot using the same axes.



Step 2

Find the number of values that will be in the box and the whiskers. Since you are making a 90% box and whisker plot, 90% of the values will be in the box and 5% of the values will be in each whisker on both sides of the box.

This box and whisker plot also will have at least 900 values in the box and no more than 50 values for each whisker.

Step 3

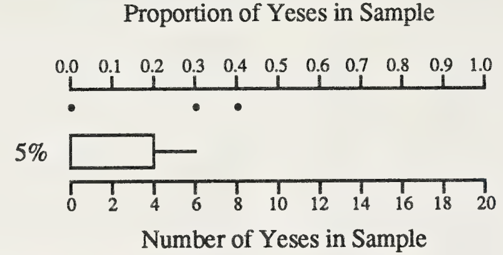
Identify the number of yeses that will mark the end of the box and the whiskers.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	126	0.126	These values represent the box.
1	0.05	214	0.214	
2	0.10	219	0.219	
3	0.15	180	0.180	
4	0.20	121	0.121	
5	0.25	87	0.087	
6	0.30	39	0.039	
7	0.35	12	0.012	These values represent the right whisker.
8	0.40	2	0.002	
9	0.45	0	0	

The ends of the box will be at 0 yeses and 6 yeses, and the right end of the right whisker will be at 8 yeses.

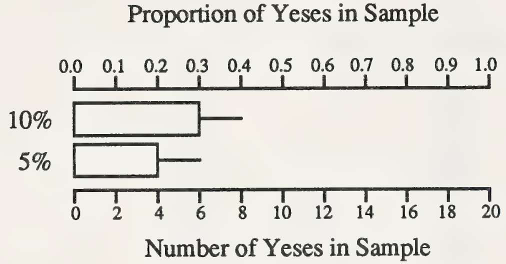
Step 4

Find the ends of the box and the whiskers on the axes. Mark these locations off with dots.



Step 5

Using the dots as your guides, draw in the box and the whiskers.



This box and whisker plot is now complete.

Draw box and whisker plots for any four of the following 90% box and whisker plots.

1. The following is for a population percentage of 15%. The sample size is 20.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	55	0.055
1	0.05	149	0.149
2	0.10	166	0.166
3	0.15	215	0.215
4	0.20	169	0.169
5	0.25	132	0.132
6	0.30	59	0.059
7	0.35	47	0.047
8	0.40	8	0.008
9	0.45	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

2. The following is for a population percentage of 20%. The sample size is 20.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	17	0.017
1	0.05	59	0.059
2	0.10	136	0.136
3	0.15	183	0.183
4	0.20	170	0.170
5	0.25	141	0.141
6	0.30	103	0.103
7	0.35	77	0.077
8	0.40	63	0.063
9	0.45	49	0.049
10	0.50	2	0.002
11	0.55	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

3. The following is for a population percentage of 25%. The sample size is 20.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	4	0.004
1	0.05	47	0.047
2	0.10	82	0.082
3	0.15	117	0.117
4	0.20	151	0.151
5	0.25	157	0.157
6	0.30	138	0.138
7	0.35	117	0.117
8	0.40	101	0.101
9	0.45	49	0.049
10	0.50	20	0.020
11	0.55	11	0.011
12	0.60	5	0.005
13	0.65	1	0.001
14	0.70	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

4. The following is for a population percentage of 30%. The sample size is 20.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	0	0
1	0.05	22	0.022
2	0.10	39	0.039
3	0.15	54	0.054
4	0.20	93	0.093
5	0.25	107	0.107
6	0.30	123	0.123
7	0.35	140	0.140
8	0.40	134	0.134
9	0.45	108	0.108
10	0.50	83	0.083
11	0.55	57	0.057
12	0.60	32	0.032
13	0.65	8	0.008
14	0.70	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

5. The following is for a population percentage of 35%. The sample size is 20.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials
0	0.00	0	0
1	0.05	7	0.007
2	0.10	52	0.052
3	0.15	78	0.078
4	0.20	102	0.102
5	0.25	111	0.111
6	0.30	123	0.123
7	0.35	129	0.129
8	0.40	113	0.113
9	0.45	98	0.098
10	0.50	76	0.076
11	0.55	57	0.057
12	0.60	42	0.042
13	0.65	11	0.011
14	0.70	0	0
15	0.75	1	0.001
16	0.80	0	0
.	.	.	.
.	.	.	.
.	.	.	.
20	1.00	0	0
Total		1000	1.000

For solutions to **Extra Help**, turn to **Appendix A, Topic 3**.

Extensions

Using a Microcomputer to Simulate Yes/No Responses

In Activity 1 you were asked to use a microcomputer to simulate the responses to a yes/no question from forty different samples. Each sample was to contain thirty responses and the population percentage was to be 43%.

Look at how this can be done using the BASIC computing language of the Apple II [1] series microcomputer (Applesoft).

- Have the microcomputer select a number between 0 and 99 inclusive.

```
140 S(A, B) = INT(100*RND (1))
```

The command RND(1) will select a random number between 0 and 1 with nine digits after the decimal point. This number is multiplied by 100, giving you a number between 0 and 100 with seven digits after the decimal point. The command INT will cut off the digits after the decimal. This will leave you with a whole number between 0 and 99 inclusive. The number will be stored in the location called S(A, B).

[1] Apple II ™ is a trademark of Apple Computers, Inc.

- Have the microcomputer select thirty numbers for each sample.

```
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
180 NEXT B
```

This FOR/NEXT loop will cause the microcomputer to select thirty different numbers and store them in the locations S(A, 1) to S(A, 30).

- Have the microcomputer select forty groups of thirty samples.

```
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
180 NEXT B
190 NEXT A
```

This FOR/NEXT loop will cause the microcomputer to select forty different groups of thirty numbers and store them in the locations S(1, 1) to S(40, 30).

- Now that the numbers have been selected, have the microcomputer decide what response each number represents, either yes or no.

```
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
180 NEXT B
190 NEXT A
```

In this particular case, if the number is less than 43, then the responses are yes. If the number of responses is greater than or equal to 43, then the responses are no. The responses are stored in the locations S$(1, 1) to S$(40, 30).

- Have the microcomputer count the number of positive responses for each sample.

```
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
```

Here the microcomputer will increase the number located at N(A) by one each time a positive response is encountered.

- Since the responses are being stored in an array, the microcomputer must be informed of the sizes of the arrays. This allows the microcomputer to set aside the memory to store the responses in these locations.

```
110 DIM N(40), S(40, 30), S$(40, 30)
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
```

The numbers that are replacing the letters in this line are informing the microcomputer of how much space is going to be required for these arrays.

- Since you do not want any errant data interfering with your work, have the microcomputer clear out all numbers that might be lingering in its memory.

```
100 CLEAR
110 DIM N(40), S(40, 30), S$(40, 30)
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
```

This command will cause all numbers in all memory locations to be set to zero.

When this program is RUN, the microcomputer will simulate all of the responses. The only remaining problem is that the microcomputer has not been told to tell you the responses.

- Tell the microcomputer to print out a response.

```
100 CLEAR
110 DIM N(40), S(40, 30), S$(40, 30)
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
230 PRINT S$(A, B);
```

The semicolon at the end of the line will cause the next response to be printed right beside the previous response.

- Tell the microcomputer to print out all thirty of the responses in a sample.

```
100 CLEAR
110 DIM N(40), S(40, 30), S$(40, 30)
120 FOR A = 1 TO 40
130 FOR B = 1 TO 30
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
220 FOR B = 1 TO 30
230 PRINT S$(A, B);
240 NEXT B
```

- Have the microcomputer print out all forty of the samples.

```
100  CLEAR
110  DIM N(40), S(40, 30), S$(40, 30)
120  FOR A = 1 TO 40
130  FOR B = 1 TO 30
140  S(A, B) = INT(100*RND (1))
150  IF S(A, B) < 43 THEN S$(A, B) = "Y"
160  IF S(A, B) >= 43 THEN S$(A, B) = "N"
170  IF S(A, B) < 43 THEN N(A) = N(A) + 1
180  NEXT B
190  NEXT A
200  FOR A = 1 TO 40
220  FOR B = 1 TO 30
230  PRINT S$(A, B);
240  NEXT B
260  NEXT A
```

- At the end of each sample, have the microcomputer print out the number of positive responses in that sample. This line also will prevent all of the responses from all of the samples from being printed together.

```
100  CLEAR
110  DIM N(40), S(40, 30), S$(40, 30)
120  FOR A = 1 TO 40
130  FOR B = 1 TO 30
140  S(A, B) = INT(100*RND (1))
150  IF S(A, B) < 43 THEN S$(A, B) = "Y"
160  IF S(A, B) >= 43 THEN S$(A, B) = "N"
170  IF S(A, B) < 43 THEN N(A) = N(A) + 1
180  NEXT B
190  NEXT A
200  FOR A = 1 TO 40
220  FOR B = 1 TO 30
230  PRINT S$(A, B);
240  NEXT B
250  PRINT " "; N(A)
260  NEXT A
```

- You can continue to modify the program by numbering each sample. This will give your presentation a cleaner look.

```
100  CLEAR
110  DIM N(40), S(40, 30), S$(40, 30)
120  FOR A = 1 TO 40
130  FOR B = 1 TO 30
140  S(A, B) = INT(100*RND (1))
150  IF S(A, B) < 43 THEN S$(A, B) = "Y"
160  IF S(A, B) >= 43 THEN S$(A, B) = "N"
170  IF S(A, B) < 43 THEN N(A) = N(A) + 1
180  NEXT B
190  NEXT A
200  FOR A = 1 TO 40
210  PRINT A;" ";
220  FOR B = 1 TO 30
230  PRINT S$(A, B);
240  NEXT B
250  PRINT " "; N(A)
260  NEXT A
```

- A printout of the responses can be obtained by using the following PR commands.

```
100  CLEAR
110  DIM N(40), S(40, 30), S$(40, 30)
120  FOR A = 1 TO 40
130  FOR B = 1 TO 30
140  S(A, B) = INT(100*RND (1))
150  IF S(A, B) < 43 THEN S$(A, B) = "Y"
160  IF S(A, B) >= 43 THEN S$(A, B) = "N"
170  IF S(A, B) < 43 THEN N(A) = N(A) + 1
180  NEXT B
190  NEXT A
195  PR#1
200  FOR A = 1 TO 40
210  PRINT A;" ";
220  FOR B = 1 TO 30
230  PRINT S$(A, B);
240  NEXT B
250  PRINT " "; N(A)
260  NEXT A
265  PR#0
```

PR#1 will send the upcoming information to the printer instead of the screen and PR#0 will send the information to the screen instead of the printer.

The program is now complete.

Answer the following questions.

1. Use the program that was created in this extension to create a new set of data. Once you have created the data, compile the information into a sampling distribution.

2. Make changes to the program that was created in this extension so the microcomputer will select sixty samples with twenty responses in each sample.

✓ For solutions to **Extensions**, turn to **Appendix A, Topic 3**.

Unit Summary

What You Have Learned

In this unit the following concepts were discussed:

- terminology relating to dispersion and probability
- techniques for developing standard deviation for both grouped and ungrouped data
- the characteristics of a normal distribution
- definition of and method for calculating *z*-scores
- definition of statistical and theoretical probabilities
- solving problems using both statistical and theoretical probabilities
- application of *z*-scores and the normal curve to solve problems involving statistical probabilities
- sets of bivariate data plotted to produce a scatterplot
- median fit method for finding the line of best fit on a scatterplot
- the equation of the line of best fit and the use of the equation for predictions
- positive and negative correlations between the variables of a bivariate distribution
- construction of 90% box and whisker plots
- the use of a 90% box and whisker plot chart to find the confidence interval for a survey result
- the confidence with which conclusions and inferences are made based on the results of yes/no surveys

You are now ready to complete the **Unit Assignment**.

Appendices

Appendix A
Solutions

Review

Topic 1 The Normal Distribution

Topic 2 Bivariate Data

Topic 3 Confidence

Appendix B
Charts and Tables

The Standard Normal Distribution Table

90% Box and Whisker Plots from Samples of Size 20

90% Box and Whisker Plots from Samples of Size 40

90% Box and Whisker Plots from Samples of Size 60

90% Box and Whisker Plots from Samples of Size 80

90% Box and Whisker Plots from Samples of Size 100

Table of Random Numbers

Appendix A
Solutions

Review

1. c the set of people or things on which information is required
   e the average of the data
   k the difference between the highest and lowest values in the data
   j measures of how the data is spread
   a facts or information
   p the distance between the lower class boundaries of two adjacent classes
   f the value that appears most often in the data
   h data that is countable
   q the number of values within each class
   n the upper and lower values between which a piece of data must fit
   o the midpoint of a class

2. a. $\text{Range} = 409 - 289$
   $= 120$

   b. $\text{Range} = 47 - 6$
   $= 41$

3. a. The mean is calculated as follows:

$$\bar{x} = \frac{a_1 + a_2 + a_3 + \ . \ . \ . \ + a_n}{n}$$
$$= \frac{346 + 325 + 295 + \ . \ . \ . \ + 349}{24}$$
$$\doteq 362.8$$

The mean is 362.8.

The median is calculated as follows:
First, put the values in ascending order.

289	295	325	336	345	346	347	349
352	354	355	356	361	366	372	376
382	386	389	398	402	409	409	409

Take the average of the middle two values, since there is an even number (24) of data values.

289	295	325	336	345	346	347	349
352	354	355	356	361	366	372	376
382	386	389	398	402	409	409	409

$$\text{median} = \frac{356 + 361}{2}$$
$$\doteq 358.5$$

The median is 358.5.

The mode is calculated as follows:

The mode is the value that appears most often.

289 295 325 336 345 346 347 349
352 354 355 356 361 366 372 376
382 386 389 398 402 409 409 409

The mode is 409.

b. The mean is calculated as follows:
First, find the products of the class marks and the frequencies. The class mark is the average of the class limits. Another name for class mark is class midpoint. Then, find the sum of the products $(f \cdot x)$ and the sum of the frequencies.

Class	Class Limits	Frequency	Class Mark	$(f \cdot x)$
1	6 - 11	5	8.5	42.5
2	12 - 17	11	14.5	159.5
3	18 - 23	23	20.5	471.5
4	24 - 29	45	26.5	1192.5
5	30 - 35	27	32.5	877.5
6	36 - 41	13	38.5	500.5
7	42 - 47	2	44.5	89
		126		3333

The mean will be the sum of the frequencies divided into the sum of the products $(f \cdot x)$.

$$\bar{x} = \frac{3333}{126}$$

$$\bar{x} \doteq 26.5$$

The median is calculated as follows:
The median is the middle term.

The cumulative frequency of the median is half the total cumulative frequency. The cumulative frequency of the median is $\frac{1}{2}(126)$ or 63. A cumulative frequency of 63 is located in the interval 24 - 29. The boundaries of this interval are 23.5 and 29.5. Class boundaries are halfway between the upper limit of one class and the lower limit of the next class.

Class	Class Limits	Class Boundaries	Frequency	Class Mark	Cumulative Frequency
1	6 - 11	5.5 - 11.5	5	8.5	5
2	12 - 17	11.5 - 17.5	11	14.5	16
3	18 - 23	17.5 - 23.5	23	20.5	39
4	24 - 29	23.5 - 29.5	45	26.5	84
5	30 - 35	29.5 - 35.5	27	32.5	111
6	36 - 41	35.5 - 41.5	13	38.5	124
7	42 - 47	41.5 - 47.5	2	44.5	126

The median then can be found by using the following procedure.

$$\frac{(\text{median} - \text{lower boundary})}{\text{upper boundary} - \text{lower boundary}} = \frac{\text{frequency of median} - \text{lower boundary cumulative frequency}}{\text{upper boundary cumulative frequency} - \text{lower boundary cumulative frequency}}$$

$$\frac{y-23.5}{29.5-23.5} = \frac{63-39}{84-39}$$

$$\frac{y-23.5}{6} = \frac{24}{45}$$

$$y = \frac{24 \times 6}{45} + 23.5$$

$$y = 26.7$$

The median of the grouped data is 26.7.

The mode is calculated as follows:

The mode is the midpoint of the class (class mark) which has the greatest frequency.

The fourth class has the greatest frequency. Therefore, the mode is 26.5.

4. Both the mean and the median are acceptable measures of the centre of the data. They are both near the centre of the data. The mode is not acceptable since it is represented by the largest values in the data.

5. a. Westcoast has 11% of the employees.

b. 42% of 15 000 000 is $0.42 \times 15\ 000\ 000 = 6\ 300\ 000$.
There are 6 300 000 people working in Centreland.

c. $$\begin{aligned}\text{Great Plains} + \text{Westcoast} + \text{Eastcoast} &= 24\% + 11\% + 7\% \\ &= 42\%\end{aligned}$$

The same percentage of people work in both of these regions.

d. $24\% - 16\% = 8\%$
8% of 15 000 000 is $0.08 \times 15\ 000\ 000 = 1\ 200\ 000$.

There are 1 200 000 more people working in the Great Plains than in Northland.

6. Step 1: Find the range of the data. The range is the lowest data value subtracted from the highest data value.

$$\begin{aligned}\text{Range} &= 409 - 287 \\ &= 122\end{aligned}$$

Step 2: Decide on the number of classes and find the class width. It is common to use between five and twenty classes. Let there be seven classes.
The class width is equal to the range divided by the number of classes. This value is rounded to the next highest whole number.

$$\begin{aligned}\text{Class width} &= \frac{122}{7} \\ &\doteq 17.4 \\ &\doteq 18 \quad \text{(rounded to the next highest whole number)}\end{aligned}$$

Step 3: Using the class width, set up the class limits and the class mark. To obtain the upper limit for the first class, add one less than the class width to the lowest data score. The upper limit is $(18 - 1) + 287 = 304$. The class limits for the first class are 287 - 304. The second class must have a lower limit of 305 which is the next integer after 304. Now 17 is added to this lower limit to obtain the upper limit. The class limits for the second class are 305 - 322. This same procedure is used to obtain the class limits for all the remaining classes. The class mark is the average of the lower and upper limits for each class. For example, the class mark for the first class equals $\frac{287+304}{2} = 295.5$.

Class	Class Limits	Class Mark
1	287 - 304	295.5
2	305 - 322	313.5
3	323 - 340	331.5
4	341 - 358	349.5
5	359 - 376	367.5
6	377 - 394	385.5
7	395 - 412	403.5

Step 4: From the given information each value can be classified into one of the seven classes by using tally marks.

Class	Class Limits	Class Mark	Tally
1	287 - 304	295.5	|||
2	305 - 322	313.5		
3	323 - 340	331.5	|||||
4	341 - 358	349.5	||||| ||||| |||||
5	359 - 376	367.5	||||| ||||| |||
6	377 - 394	385.5	|||||
7	395 - 412	403.5	|||||	

Step 5: The frequency for each class is the tally total for each class. Complete the frequency column and note that the sum of this frequency column should equal the number of data values. For this problem, there are forty-eight given data values.

Class	Class Limits	Class Mark	Tally	Frequency
1	287 - 304	295.5	|||	3
2	305 - 322	313.5			1
3	323 - 340	331.5	|||||	5
4	341 - 358	349.5	||||| ||||| |||||	15
5	359 - 376	367.5	||||| ||||| |||	13
6	377 - 394	385.5	|||||	5
7	395 - 412	403.5	|||||		6

7. a. $\bar{x}=\dfrac{a_1+a_2+a_3+\ .\ .\ .\ +a_n}{n}$

$\bar{x}=\dfrac{18+34+14}{3}$

$\bar{x}=22$

b. $\bar{x}=\dfrac{a_1+a_2+a_3+\ .\ .\ .\ +a_n}{n}$

$\bar{x}=\dfrac{6+4+7+10+9+9+11+16}{8}$

$\bar{x}=9$

c. $\bar{x} = \dfrac{a_1 + a_2 + a_3 + \ . \ . \ . \ + a_n}{n}$

$\bar{x} = \dfrac{141+107+118+114+125}{5}$

$\bar{x} = 121$

8. a. First, arrange the values in ascending order.
14, 18, 34
Now, select the middle value or take the average of the two middle values.
18

b. First, arrange the values in ascending order.
4, 6, 7, 9, 9, 10, 11, 16
Now, select the middle value or take the average of the two middle values.

$\dfrac{9+9}{2} = \dfrac{18}{2}$

$= 9$

c. First, arrange the values in ascending order.
107, 114, 118, 125, 141
Now, select the middle value or take the average of the two middle values.
118

9. a. $m = \dfrac{y_1 - y_2}{x_1 - x_2}$

$m = \dfrac{7-5}{4-9}$

$m = \dfrac{2}{-5}$

$m = \dfrac{-2}{5}$

b. $m = \dfrac{y_1 - y_2}{x_1 - x_2}$

$m = \dfrac{22-102}{108-112}$

$m = \dfrac{-80}{-4}$

$m = 20$

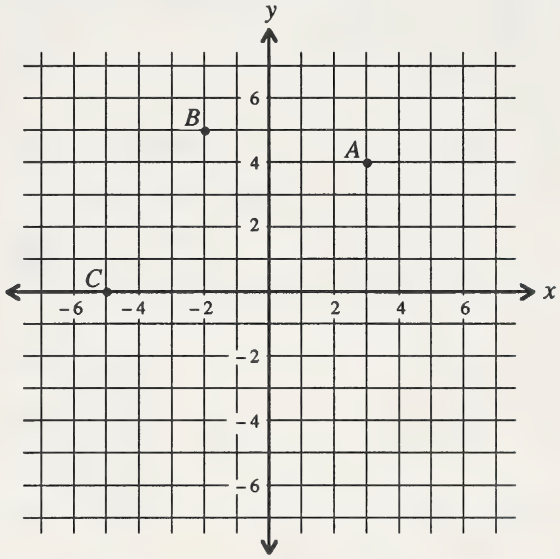
10. a.

$$\begin{aligned} y - y_1 &= m(x - x_1) \\ y - 7 &= \frac{-2}{5}(x-4) \\ 5y - 35 &= -2x + 8 \\ 2x + 5y - 43 &= 0 \end{aligned}$$

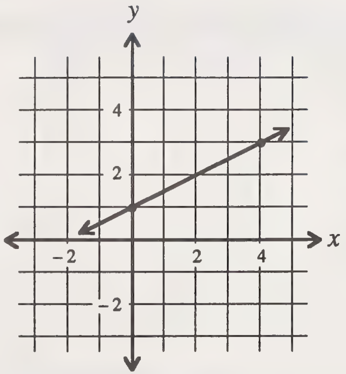
b.

$$\begin{aligned} y - y_1 &= m(x - x_1) \\ y - 22 &= 20(x - 108) \\ y - 22 &= 20x - 2160 \\ 20x - y - 2138 &= 0 \end{aligned}$$

11. a. $A(3, 4)$ b. $B(-2, 5)$ c. $C(-5, 0)$



12. a.



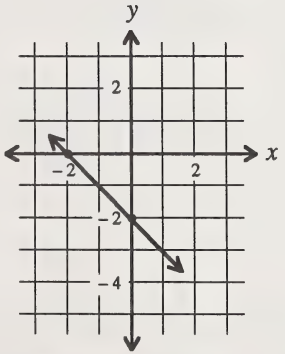
The first point is (0, 1).
The second point is (4, 3).

$$m=\frac{y_1-y_2}{x_1-x_2}$$

$$m=\frac{1-3}{0-4}$$

$$m=\frac{1}{2}$$

b.



The first point is (– 2, 0).
The second point is (0, – 2).

$$m=\frac{y_1-y_2}{x_1-x_2}$$

$$m=\frac{0-(-2)}{(-2)-0}$$

$$m=-1$$

13. a. $\frac{12}{25}$

$$\frac{12}{25}\times 100\% = 0.48\times 100\%$$
$$= 48\%$$

b. $\frac{17}{20}$

$$\frac{17}{20}\times 100\% = 0.85\times 100\%$$
$$= 85\%$$

c. $\frac{9}{60}$

$$\frac{9}{60}\times 100\% = 0.15\times 100\%$$
$$= 15\%$$

Exploring Topic 1

Activity 1

Calculate and interpret the mean and the standard deviation of a set of data.

1. Dispersion is a measure that describes how data is spread.

2. a. $\mu = \dfrac{102 + 97 + 105 + \ . \ . \ . \ + 98}{12}$

$\mu = \dfrac{1224}{12}$

$\mu = 102$

x	μ	$x - \mu$	$(x-\mu)^2$
102	102	0	0
97	102	−5	25
105	102	3	9
105	102	3	9
96	102	−6	36
111	102	9	81
106	102	4	16
100	102	−2	4
99	102	−3	9
103	102	1	1
102	102	0	0
98	102	−4	16
		0	206

b. $\sum (x-\mu) = 0$

c. $\sum (x-\mu)^2 = 206$

d. $\sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{n}}$

$\sigma = \sqrt{\dfrac{206}{12}}$

$\sigma = \sqrt{17.17}$

$\sigma \doteq 4.1$

3.

New Duetch

x	μ	$x-\mu$	$(x-\mu)^2$
198	200	-2	4
205	200	5	25
201	200	1	1
200	200	0	0
207	200	7	49
192	200	-8	64
194	200	-6	36
199	200	-1	1
206	200	6	36
210	200	10	100
198	200	-2	4
193	200	-7	49
195	200	-5	25
192	200	-8	64
207	200	7	49
203	200	3	9
			516

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{516}{16}}$$

$$\sigma = \sqrt{32.25}$$

$$\sigma \doteq 5.7$$

Hosts

x	μ	$x-\mu$	$(x-\mu)^2$
199	200	-1	1
201	200	1	1
201	200	1	1
197	200	-3	9
201	200	1	1
198	200	-2	4
196	200	-4	16
203	200	3	9
198	200	-2	4
206	200	6	36
203	200	3	9
196	200	-4	16
206	200	6	36
197	200	-3	9
202	200	2	4
196	200	-4	16
			172

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{172}{16}}$$

$$\sigma = \sqrt{10.75}$$

$$\sigma \doteq 3.3$$

The Hosts company is more consistent in its packaging.

4. All the pieces in the set of data would have the same value.

5.

Class	Number Sold	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	1 - 3	2	2	4	-9	81	162
2	4 - 6	3	5	15	-6	36	108
3	7 - 9	5	8	40	-3	9	45
4	10 - 12	10	11	110	0	0	0
5	13 - 15	6	14	84	3	9	54
6	16 - 18	3	17	51	6	36	108
7	19 - 21	2	20	40	9	81	162
		31		344	0	252	639

a. $\mu = \dfrac{\sum (f \cdot x)}{n}$

$\mu = \dfrac{344}{31}$

$\mu \doteq 11$ (to the nearest whole number)

b. $\sum (x-\mu) = 0$

c. $\sum (x-\mu)^2 = 252$

d. $\sigma = \sqrt{\dfrac{\sum f(x-\mu)^2}{n}}$

$\sigma = \sqrt{\dfrac{639}{31}}$

$\sigma \doteq 4.5$

6. a.

The Big Stone

Class	Percent Increase	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	0 - 0.99	3	0.495	1.485	– 2.367	5.603	16.809
2	1 - 1.99	5	1.495	7.475	– 1.367	1.869	9.345
3	2 - 2.99	9	2.495	22.455	– 0.367	0.135	1.215
4	3 - 3.99	7	3.495	24.465	0.633	0.401	2.807
5	4 - 4.99	3	4.495	13.485	1.633	2.667	8.001
6	5 - 5.99	3	5.495	16.485	2.633	6.933	20.799
		30		85.850			58.976

$$\mu = \frac{\sum (f \cdot x)}{n}$$

$$\mu = \frac{85.850}{30}$$

$$\mu \doteq 2.862$$

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{58.976}{30}}$$

$$\sigma \doteq 1.4$$

Royal Eastern

Class	Percent Increase	Frequency (f)	Class Midpoint (x)	Product $(f \cdot x)$	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
1	0 - 0.99	3	0.495	1.485	– 2.4	5.76	17.28
2	1 - 1.99	6	1.495	8.97	– 1.4	1.96	11.76
3	2 - 2.99	7	2.495	17.465	– 0.4	0.16	1.12
4	3 - 3.99	7	3.495	24.465	0.6	0.36	2.52
5	4 - 4.99	4	4.495	17.98	1.6	2.56	10.24
6	5 - 5.99	3	5.495	16.485	2.6	6.76	20.28
		30		86.850			63.20

$$\mu = \frac{\sum(f \cdot x)}{n}$$

$$\mu = \frac{86.85}{30}$$

$$\mu = 2.895$$

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{63.2}{30}}$$

$$\sigma \doteq 1.5$$

b. The Big Stone has the most consistent increases in premiums because the standard deviation is smaller, and thus, the dispersion around the mean is less.

c. Royal Eastern has the largest sum for the $f(x-\mu)^2$ column.

Activity 2

Identify the normal distribution.

1.
$$\begin{aligned} 25 &= \mu + 2\sigma \quad ① \\ \underline{10} &\underline{= \mu - 1\sigma} \quad ② \\ 15 &= 3\sigma \quad ① - ② \\ 5 &= \sigma \end{aligned}$$

$$\begin{aligned} 10 &= \mu - 1\sigma \\ 10 &= \mu - 1(5) \\ 10 &= \mu - 5 \\ 15 &= \mu \end{aligned}$$

$$\begin{aligned} a &= \mu - 3\sigma \\ a &= 15 - 3(5) \\ a &= 15 - 15 \\ a &= 0 \end{aligned}$$

$$\begin{aligned} b &= \mu - 2\sigma \\ b &= 15 - 2(5) \\ b &= 15 - 10 \\ b &= 5 \end{aligned}$$

$$\begin{aligned} c &= \mu \\ c &= 15 \end{aligned}$$

$$\begin{aligned} d &= \mu + 1\sigma \\ d &= 15 + 1(5) \\ d &= 15 + 5 \\ d &= 20 \end{aligned}$$

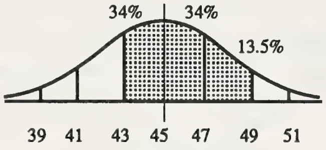
$$\begin{aligned} e &= \mu + 3\sigma \\ e &= 15 + 3(5) \\ e &= 15 + 15 \\ e &= 30 \end{aligned}$$

2. a. Since 2 g is one standard deviation, the question asks how many of the chocolates are within one standard deviation of the mean. 68% of the 10 000 chocolates will be within one standard deviation of the mean.

$$\begin{aligned} 68\% \text{ of } 10\,000 \text{ is } 68\% \times 10\,000 &= 0.68 \times 10\,000 \\ &= 6800. \end{aligned}$$

There will be 6800 chocolates within 2 g of the required mass.

b. First, construct the graph of the situation.

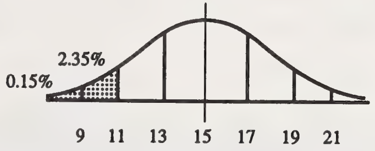


The percent rejected will be 100% – (34% + 34% + 13.5%) or 18.5%.

$$18.5\% \text{ of } 100\ 000 \text{ is } 18.5\% \times 100\ 000 = 0.185 \times 100\ 000$$
$$= 18\ 500.$$

There will be 18 500 chocolates rejected.

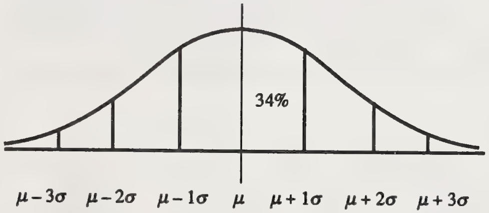
3. First, construct the graph using this information.



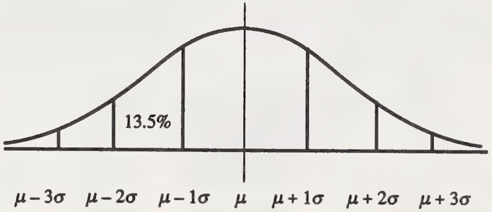
2.5% is below two standard deviations less than the mean. Two standard deviations less than the mean is eleven years.

The company will offer a guarantee of eleven years.

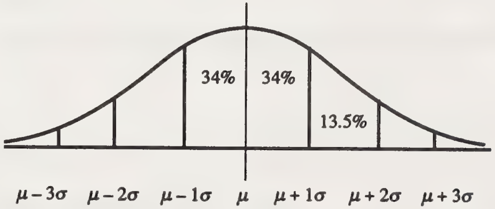
4. a. The area between μ and $\mu + \sigma$ is 34%.



b. The area between $\mu - 2\sigma$ and $\mu - \sigma$ is 13.5%.

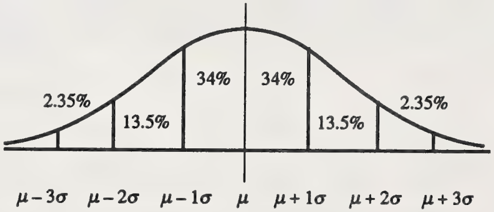


c. The area between $\mu + 2\sigma$ and $\mu - \sigma$ is 81.5%.



34% + 34% + 13.5% = 81.5%

d. The area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.7%.



2.35% + 13.5% + 34% + 34% + 13.5% + 2.35% = 99.7%

Activity 3

Use *z*-scores to solve situations that are normally distributed.

1. The *z*-score is -2.

2. The mean is 197.
   The standard deviation is approximately 5.6.
   The *z*-scores are as follows:

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{188-197}{5.6} \\ &= \frac{-9}{5.6} \\ &\doteq -1.61 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{190-197}{5.6} \\ &= \frac{-7}{5.6} \\ &= -1.25 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{192-197}{5.6} \\ &= \frac{-5}{5.6} \\ &\doteq -0.89 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{196-197}{5.6} \\ &= \frac{-1}{5.6} \\ &\doteq -0.18 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{198-197}{5.6} \\ &= \frac{1}{5.6} \\ &\doteq 0.18 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{199-197}{5.6} \\ &= \frac{2}{5.6} \\ &\doteq 0.36 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{202-197}{5.6} \\ &= \frac{5}{5.6} \\ &\doteq 0.89 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{207-197}{5.6} \\ &= \frac{10}{5.6} \\ &\doteq 1.79 \end{aligned}$$

3. Find the *z*-score for each half of the round.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{40-43}{2} \\ &= \frac{-3}{2} \\ &= -1.5 \end{aligned}$$

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{38-41}{1.8} \\ &= \frac{-3}{1.8} \\ &\doteq -1.67 \end{aligned}$$

The half round that has the lower *z*-score will be the better round.

Andrea had a better round on the back nine.

4. The *z*-score for 185 hits is as follows:

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{185-150}{20} \\ &= \frac{35}{20} \\ &= 1.75 \end{aligned}$$

The area for the *z*-score 1.75 is 0.4599.

The following represents the area to the right of the *z*-score 1.75.

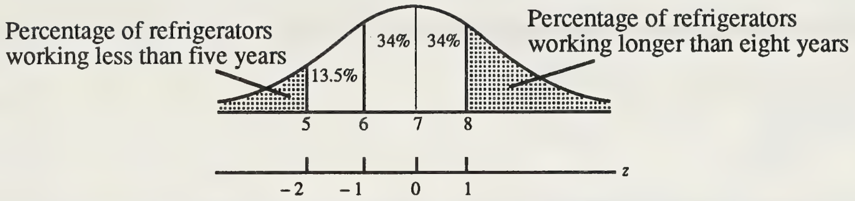
$0.5 - 0.4599 = 0.0401$

The following represents the number of seasons the player had at least 185 hits.

$$\begin{aligned} 0.0401 \text{ of } 25 \text{ is } 0.0401 \times 25 &= 1.0025 \\ &\doteq 1. \end{aligned}$$

The baseball player had at least 185 hits in only one season.

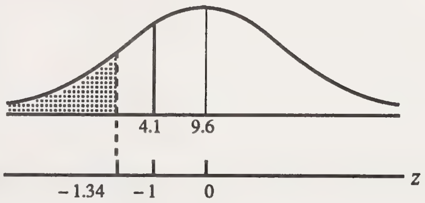
5.


Percentage of refrigerators working less than five years
Percentage of refrigerators working longer than eight years
13.5%
34%
34%
5
6
7
8
−2
−1
0
1
z

Approximate Area Method	z-score Method
The percentage of refrigerators working less than five years is 50% – (34% + 13.5%) = 2.5%. $\frac{2.5}{100}\times 1200=30$ Thirty refrigerators must be replaced. The percentage of refrigerators working more than eight years is 50% – 34% = 16%. $\frac{16}{100}\times 1200=192$ The number of refrigerators working longer than eight years is 192.	The area to the left of five years must be determined using the z-score formula and the table of values. $z=\frac{x-\mu}{\sigma}=\frac{5-7}{1}=-2$ (This matches the diagram.) From the table, the area for $z=-2$ is 0.4772. This is the area to the left of the mean between $z=0$ and $z=-2$. The required area which represents the percentage of refrigerators working less than five years is 0.5000 – 0.4772 = 0.0228. Thus, $0.0228\times 1200=27.36$. The number of refrigerators that must be replaced is 27. To determine the number of refrigerators working longer than eight years, the area to the right of eight years must be calculated. $z=\frac{8-7}{1}=1$ (This matches the diagram.) From the tables when $z=1$, the area is 0.3413. This is the area to the right of the mean between $z=0$ and $z=1$. The required area to the right of eight years is 0.5000 – 0.3413 = 0.1587. Now, $0.1587\times 1200=190.44$. Thus, 190 refrigerators will be working longer than eight years.

The z-score method is the preferred method.

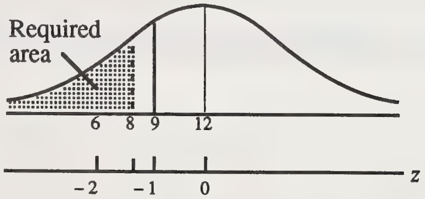
6. Since the company will only replace 9% of the toasters, then $50\% - 9\% = 41\%$ is the area to the left of the mean. From the z-score table an area of 41% to the left of the mean gives a z-score value of -1.34. The value of x is determined by using the z-score formula.



$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ -1.34 &= \frac{x-9.6}{5.5} \\ -7.37 &= x-9.6 \\ x &= 2.23 \end{aligned}$$

The guarantee period for the toasters should be 2.2 years.

7. Determine the z-score for eight years by using the formula.



$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ z &= \frac{8-12}{3} \\ z &= -1.33\dot{3} \end{aligned}$$

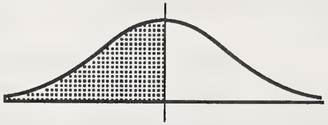
From the table, the area for this z-score is 0.4082. This is the area from the mean to $z = -1.33\dot{3}$. The required area for the period less than eight years is $0.5000 - 0.4082 = 0.0918$.

Thus, 9% of the compressors have to be repaired under this eight-year warranty period.

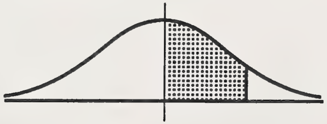
Activity 4

Use z-scores to calculate the probability of an event happening.

1. a. $P(z<0)=0.5$

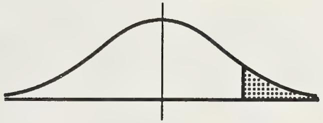


b. $P(0<z<2)=0.4772$

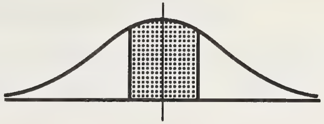


c. The area for a z-score of 1.80 is 0.4641.

$$P(z>1.80)=0.5-0.4641$$
$$=0.0359$$



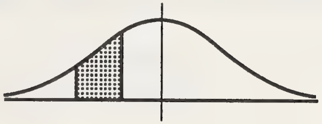
d.



The area for a z-score of -0.89 is 0.3133.
The area for a z-score of 1.24 is 0.3925.

$$P(-0.89<z<1.24)=0.3133+0.3925$$
$$=0.7058$$

e.

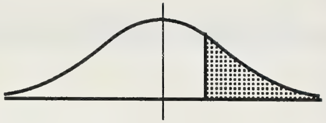


The area for the z-score -2.37 is 0.4911.
The area for the z-score -0.92 is 0.3212.

$$P(-2.37<z<-0.92)=0.4911-0.3212$$
$$=0.1699$$

2. a. $z=\dfrac{x-\mu}{\sigma}$

$$z=\frac{15-13}{1.3}$$
$$z \doteq 1.54$$



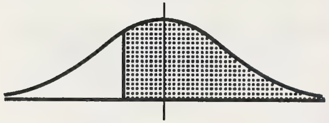
The area for the z-score 1.54 is 0.4382.

$$P(>15\ years)=0.5-0.4382$$
$$=0.0618$$

The probability that a cat will live longer than fifteen years is 0.0618.

b. $z=\dfrac{x-\mu}{\sigma}$

$$z=\frac{12-13}{1.3}$$
$$z \doteq -0.77$$



The area for the z-score -0.77 is 0.2794.

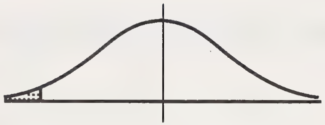
$$P(>12\ years)=0.5+0.2794$$
$$=0.7794$$

The probability that a cat will live longer than twelve years is 0.7794.

c. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{10-13}{1.3}$

$z\doteq-2.31$



The area for the z-score -2.31 is 0.4896.

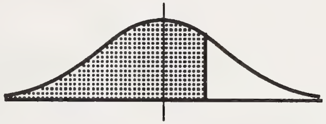
$$P(<10\ \mathit{years})=0.5-0.4896$$
$$=0.0104$$

The probability that a cat will live less than ten years is 0.0104.

d. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{14-13}{1.3}$

$z\doteq0.77$



The area for the z-score 0.77 is 0.2794.

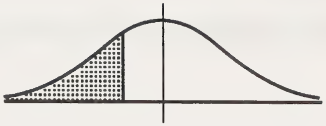
$$P(<14\ \mathit{years})=0.5+0.2794$$
$$=0.7794$$

The probability that a cat will live less than fourteen years is 0.7794.

3. a. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{32-35}{6.5}$

$z=-0.46$



The area for the z-score -0.46 is 0.1772.

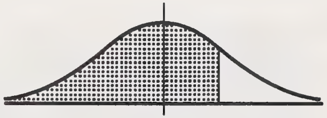
$$P(<32\ \mathit{hours})=0.5-0.1772$$
$$=0.3228$$

The probability that a battery will last less than 32 h is 0.3228.

b. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{43-35}{6.5}$

$z\doteq1.23$



The area for the z-score 1.23 is 0.3907.

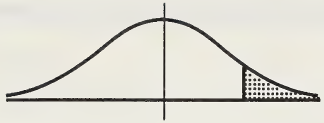
$$P(<43\ \mathit{hours})=0.5+0.3907$$
$$=0.8907$$

The probability that a battery will last less than 43 h is 0.8907.

c. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{47-35}{6.5}$

$z \doteq 1.85$



The area for the z-score 1.85 is 0.4678.

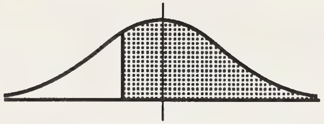
$$\begin{aligned}P(>47\ hours)&=0.5-0.4678\\&=0.0322\end{aligned}$$

The probability that a battery will last more than 47 h is 0.0322.

d. $z=\dfrac{x-\mu}{\sigma}$

$z=\dfrac{33-35}{6.5}$

$z \doteq -0.31$



The area for the z-score -0.31 is 0.1217.

$$\begin{aligned}P(>33\ hours)&=0.5+0.1217\\&=0.6217\end{aligned}$$

The probability that a battery will last more than 33 h is 0.6217.

4. First, find the z-score to go with the area 0.4826, which is -2.11.

Fill this information into the z-score formula and solve for the unknown.

$$z=\frac{x-\mu}{\sigma}$$

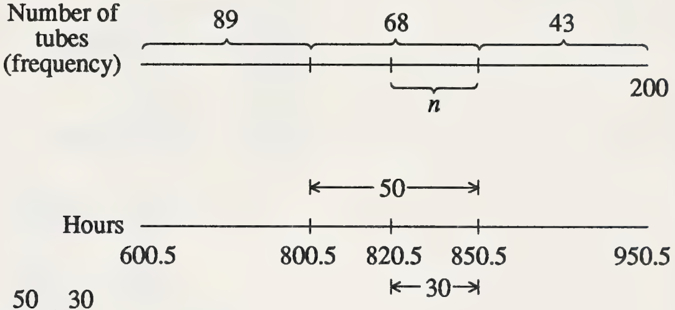
$$-2.11=\frac{1.3-3}{\sigma}$$

$$\sigma=\frac{1.3-3}{-2.11}$$

$$\sigma=\frac{-1.7}{-2.11}$$

$$\sigma \doteq 0.81$$

The standard deviation for the shelf period is 0.81 days.

5. 

$$\frac{50}{68}=\frac{30}{n}$$

$$n=\frac{30\times 68}{50}$$

$$n=40.8$$

The number of tubes lasting longer than 820.5 h is $40.8 + 27 + 16 = 83.8$.

$$\text{Probability} = \frac{\text{number of tubes that last longer than 820.5 h}}{\text{total number of tubes}}$$
$$= \frac{83.8}{200}$$
$$= 0.419$$
$$= 0.42$$

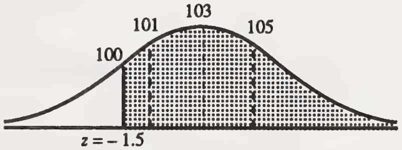
The probability a tube will last longer than 820.5 h is 0.42.

6. $z = \frac{x - \mu}{\sigma}$

$z = \frac{100 - 103}{2}$

$z = \frac{-3}{2}$

$z = -1.5$



The area for $z = -1.5$ is 0.4332.
The probability that a box will contain 100 or more tacks is $0.5 + 0.4332 = 0.9332$.

Extra Help

1. a. Type 6
   b. Type 3
   c. Type 2
   d. Types 6 and 7
   e. Type 5
   f. Type 7

2. a. $0 < z < 2.76$
   Type 1
   b. $-2.82 < z < -1.09$
   Type 5
   c. $z < -2.3$
   Type 7
   d. $z < -2$ and $z > 2$
   Types 6 and 7
   e. $2.3 < z < 2.32$
   Type 4
   f. $-0.04 < z < 1.03$
   Type 3

3. a. Area = 0.3888

   b. Area = 0.5 − 0.4971
   = 0.0029

   c. Area = 0.4066 + 0.5
   = 0.9066

   d. Left area = 0.5 − 0.4429
   = 0.0571

   Right area = 0.5 − 0.4909
   = 0.0091

   Total area = 0.0571 + 0.0091
   = 0.0662

4. a. Area = 0.5 − 0.3461
   = 0.1539

   b. Area = 0.5 + 0.4793
   = 0.9793

   c. Area = 0.4929

   d. Area = 0.4913 − 0.4706
   = 0.0207

Extensions

1. $\boxed{\text{MODE}}$ $\boxed{\cdot}$

$\boxed{\text{INV}}$ $\boxed{\text{SAC}}$

16 $\boxed{\times}$ 4 $\boxed{x}$

25 $\boxed{\times}$ 12 $\boxed{x}$

34 $\boxed{\times}$ 32 $\boxed{x}$

43 $\boxed{\times}$ 76 $\boxed{x}$

52 $\boxed{\times}$ 59 $\boxed{x}$

61 $\boxed{\times}$ 18 $\boxed{x}$

70 $\boxed{\times}$ 11 $\boxed{x}$

79 $\boxed{\times}$ 3 $\boxed{x}$

$\boxed{\text{INV}}$ $\boxed{\bar{x}}$

$\boxed{\text{INV}}$ $\boxed{\sigma n}$

The mean is approximately 46.0. The standard deviation is approximately 11.8.

2. $\boxed{\text{MODE}}$ $\boxed{\cdot}$

$\boxed{\text{INV}}$ $\boxed{\text{SAC}}$

24 $\boxed{\times}$ 8 $\boxed{x}$

29 $\boxed{\times}$ 10 $\boxed{x}$

34 $\boxed{\times}$ 29 $\boxed{x}$

39 $\boxed{\times}$ 57 $\boxed{x}$

44 $\boxed{\times}$ 73 $\boxed{x}$

49 $\boxed{\times}$ 43 $\boxed{x}$

54 $\boxed{\times}$ 22 $\boxed{x}$

59 $\boxed{\times}$ 9 $\boxed{x}$

$\boxed{\text{INV}}$ $\boxed{\bar{x}}$

$\boxed{\text{INV}}$ $\boxed{\sigma n}$

The mean is approximately 42.7. The standard deviation is approximately 7.6.

Exploring Topic 2

Activity 1

Plot sets of bivariate data to produce a scatterplot.

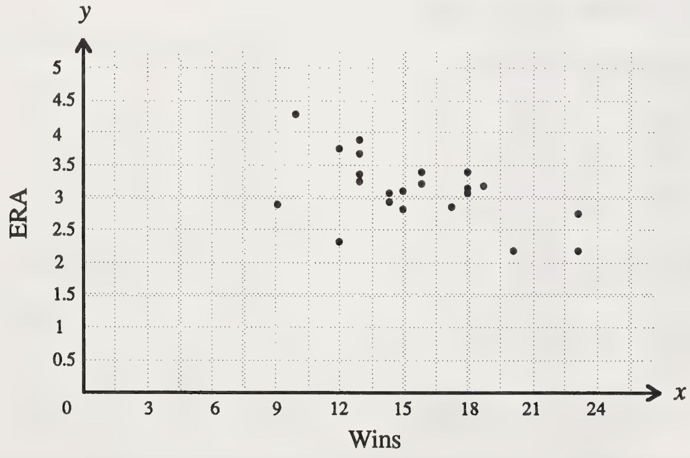
1. a. no relation

   b. Volume is independent and mass is dependent.

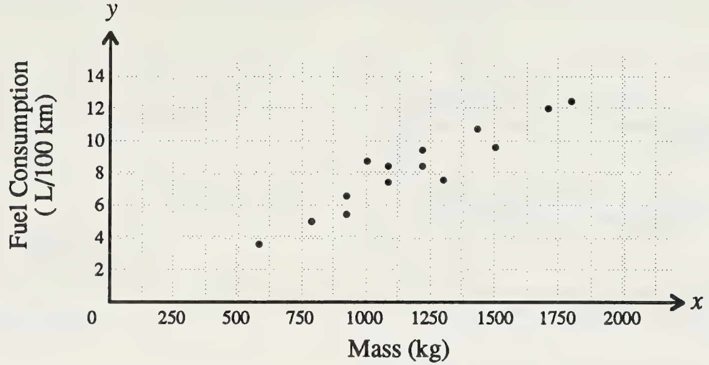
   c. no relation

   d. Time is independent and distance travelled is dependent.

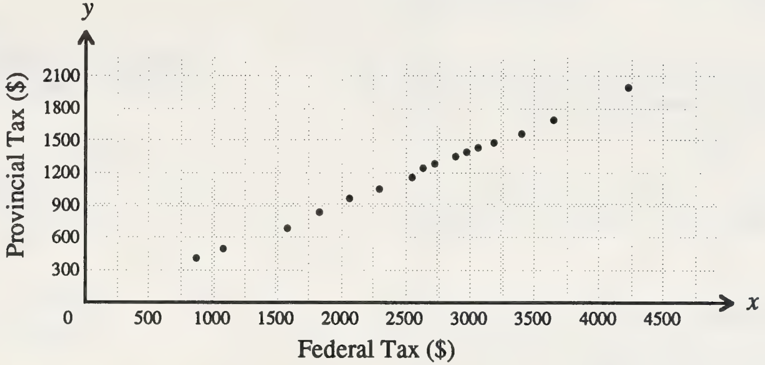
2. Wins versus ERA for Distance League Starting Pitchers, 1989



3. Mass versus Fuel Consumption of Vehicles



4. Federal and Provincial Income Tax Payable for Northland Before Surtax, 1989



Activity 2

Plot a line of best fit on a scatterplot using the median fit method.

1. There are twenty-two points; thus, there will be seven points in each of the first and third strips. The second strip will have eight points.

The first strip has the following points:
(9, 2.92), (10, 4.29), (12, 2.44), (12, 3.69), (13, 3.25), (13, 3.32), and (13, 3.67)

The median for the *x*-coordinate is as follows:
9, 10, 12, 12, 13, 13, 13
→ 12

The median for the *y*-coordinate is as follows:
2.44, 2.92, 3.25, 3.32, 3.67, 3.69, 4.29
→ 3.32

The second strip has the following points:
(13, 3.86), (14, 2.92), (14, 3.14), (15, 2.72), (15, 3.08), (16, 3.26), (16, 3.43), and (17, 2.91)

The median for the *x*-coordinate is as follows:
$13, 14, 14, \underline{15, 15}, 16, 16, 17$

$$\rightarrow \frac{15+15}{2} = 15$$

The median for the *y*-coordinate is as follows:
$2.72, 2.91, 2.92, \underline{3.08, 3.14}, 3.26, 3.43, 3.86$

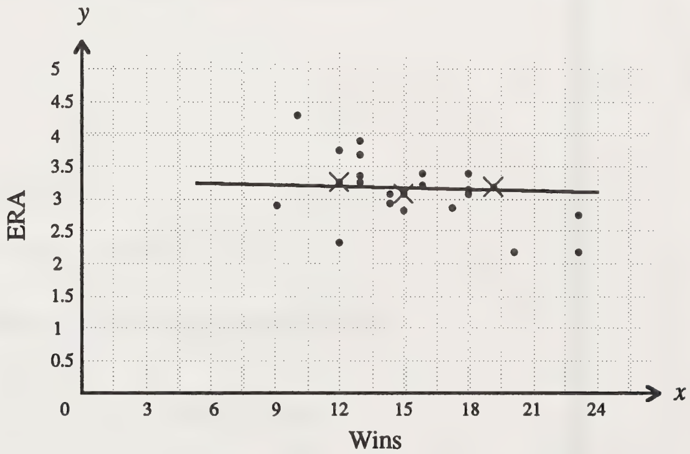
$$\rightarrow \frac{3.08+3.14}{2} = 3.11$$

The third strip has the following points:
(18, 3.18), (18, 3.19), (18, 3.41), (19, 3.12), (20, 2.22), (23, 2.26), and (23, 2.73)

The median for the *x*-coordinate is as follows:
18, 18, 18, 19, 20, 23, 23
→ 19

The median for the *y*-coordinate is as follows:
2.22, 2.26, 2.73, 3.12, 3.18, 3.19, 3.41
→ 3.12

Wins versus ERA for Distance League Starting Pitchers, 1989



2. The first strip has the following points:
(600, 3.9), (800, 5.0), (900, 5.6), (900, 6.2), and (1000, 8.5)

The median for the *x*-coordinate is as follows:
600, 800, 900, 900, 1000
→ 900

The median for the y-coordinate is as follows:
3.9, 5.0, 5.6, 6.2, 8.5

$\longrightarrow 5.6$

The second strip has the following points:
(1100, 7.5), (1100, 8.2), (1200, 8.2), and (1200, 9.6)

The median for the x-coordinate is as follows:
$1100, \underline{1100, 1200}, 1200$

$\longrightarrow \dfrac{1100+1200}{2}=1150$

The median for the y-coordinate is as follows:
$7.5, \underline{8.2, 8.2}, 9.6 \quad \longrightarrow \dfrac{8.2+8.2}{2}=8.2$

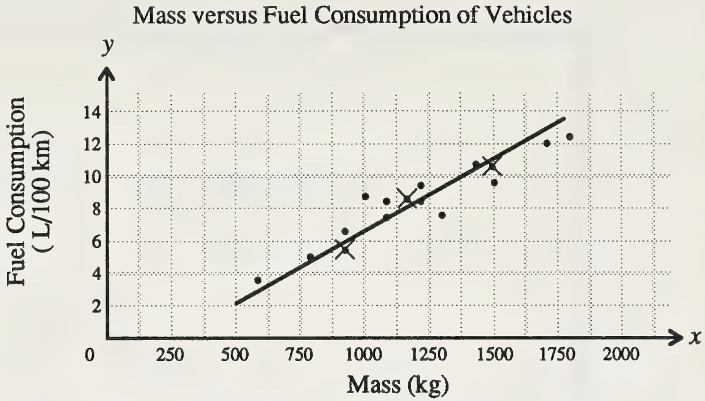
The third strip has the following points:
(1300, 7.8), (1400, 10.4), (1500, 9.8), (1700, 12.0), and (1800, 12.3)

The median for the x-coordinate is as follows:
1300, 1400, 1500, 1700, 1800

$\longrightarrow 1500$

The median for the y-coordinate is as follows:
7.8, 9.8, 10.4, 12.0, 12.3

$\longrightarrow 10.4$



3. The first strip has the following points:
(820, 381.80), (1060, 493.60), (1540, 716.40), (1780, 828.80), and (2040, 949.90)

The median for the x-coordinate is as follows:
820, 1060, 1540, 1780, 2040

$\longrightarrow 1540$

The median for the y-coordinate is as follows:
381.80, 493.60, 716.40, 828.80, 949.90

$\longrightarrow 716.40$

The second strip has the following points:
(2280, 1060.30), (2560, 1190.10), (2620, 1218.20), (2740, 1274.40), (2860, 1330.60), and (2980, 1386.80)

The median for the x-coordinate is as follows:
$2280, 2560, \underline{2620, 2740}, 2860, 2980$

$\longrightarrow \dfrac{2620+2740}{2}=2680$

The median for the y-coordinate is as follows:
1060.30, 1190.10, $\underline{1218.20, 1274.40}$, 1330.60, 1386.80

$$\longrightarrow \frac{1218.20+1274.40}{2} = 1246.30$$

The third strip has the following points:
(3020, 1404.20), (3160, 1469.10), (3340, 1553.40),
(3660, 1702.60), and (4240, 1972.90)

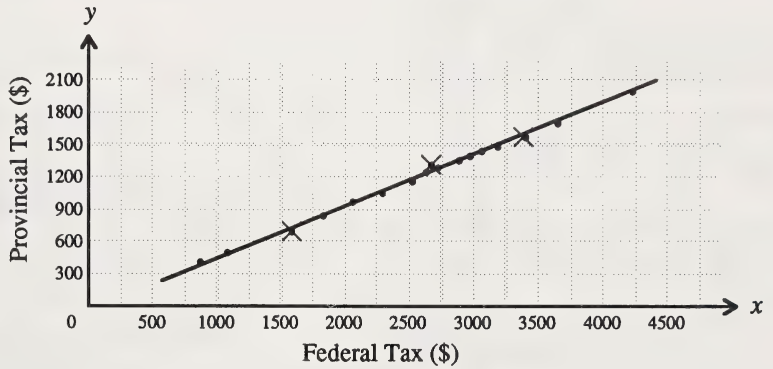
The median for the x-coordinate is as follows:
3020, 3160, 3340, 3660, 4240

$$\longrightarrow 3340$$

The median for the y-coordinate is as follows:
1404.20, 1469.10, 1553.40, 1702.60, 1972.90

$$\longrightarrow 1553.40$$

Federal and Provincial Income Tax Payable in Northland Before Surtax, 1989



Activity 3

Use the equation of the line of best fit to generate new data for a population.

1. a. i. $\doteq 4.5$ L/100 km

ii. $\doteq 9.0$ L/100 km

iii. $\doteq 13.0$ L/100 km

To answer the next two questions, first you must find the equation of the line of best fit.

$$m=\frac{y_1-y_2}{x_1-x_2}$$
$$m=\frac{13.0-4.5}{1750-750}$$
$$m=\frac{8.5}{1000} \text{ or } 0.0085$$

$$y=mx+b$$
$$y=0.0085x+b$$
$$9.0=0.0085(1275)+b$$
$$9.0=10.8375+b$$
$$-1.8375=b$$
$$y=0.0085x-1.8375$$

iv. $y=0.0085x-1.8375$

$y=0.0085(400)-1.8375$

$y=3.4-1.8375$

$y=1.5625$

1.56 L/100 km

v. $y=0.0085x-1.8375$

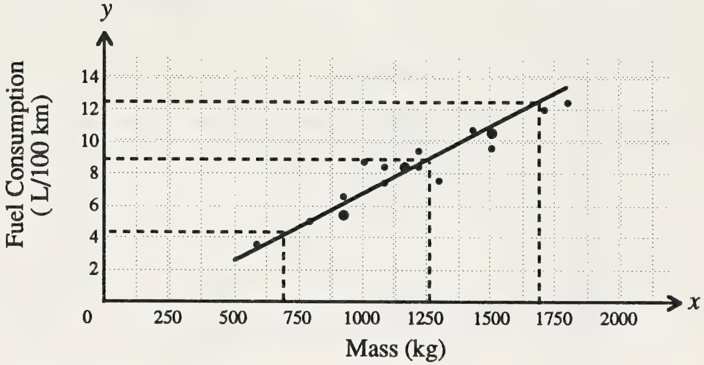
$y=0.0085(4500)-1.8375$

$y=38.25-1.8375$

$y=36.4125$

36.41 L/100 km

Mass versus Fuel Consumption of Vehicles



b. i. $\doteq 700$ kg

ii. $\doteq 1270$ kg

iii. $\doteq 1700$ kg

iv. $y=0.0085x-1.8375$

$1.2=0.0085x-1.8375$

$3.0375=0.0085x$

$357.4 \doteq x$

357 kg

v. $y = 0.0085x - 1.8375$

$17.9 = 0.0085x - 1.8375$

$19.7375 = 0.0085x$

$2322.1 \doteq x$

2322 kg

2. a. i. 8 wins

ii. 13 wins

iii. 18 wins

To answer the next two questions, you must first find the equation of the line of best fit.

$$m = \frac{y_1 - y_2}{x_1 - x_2}$$

$$m = \frac{3.40 - 3.00}{8 - 18}$$

$$m = \frac{-0.40}{10} \text{ or } -0.04$$

$$y = mx + b$$

$$y = -0.04x + b$$

$$3.2 = -0.04(13) + b$$

$$3.2 = -0.52 + b$$

$$3.72 = b$$

$$y = -0.04x + 3.72$$

iv. $y = -0.04x + 3.72$

$3.60 = -0.04x + 3.72$

$-0.12 = -0.04x$

$3 = x$

3 wins

v. $y = -0.04x + 3.72$

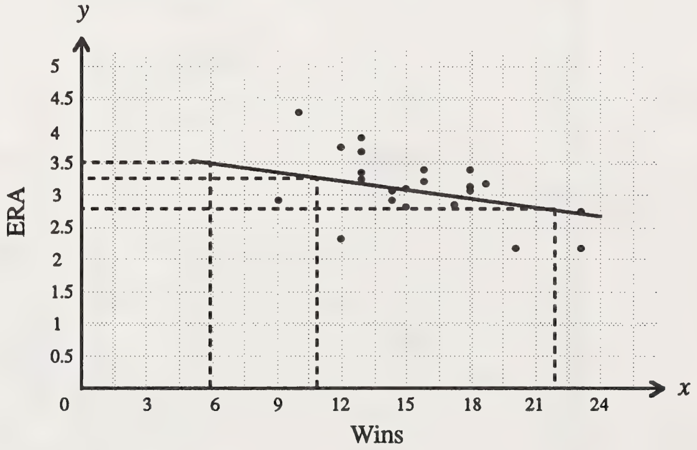
$1.90 = -0.04x + 3.72$

$-1.82 = -0.04x$

$45.5 = x$

46 wins

Wins versus ERA for Distance League Starting Pitchers, 1989



b. i. $\doteq 3.50$

ii. $\doteq 3.30$

iii. $\doteq 2.75$

iv. $y = -0.04x + 3.72$

$y = -0.04(1) + 3.72$

$y = -0.04 + 3.72$

$y = 3.68$

v. $y = -0.04x + 3.72$

$y = -0.04(30) + 3.72$

$y = -1.2 + 3.72$

$y = 2.52$

3. a. i. $\doteq \$1350$

ii. $\doteq \$2000$

iii. $\doteq \$3950$

To answer the next two questions, first you must find the equation of the line of best fit.

$$(y - y_1) = \frac{y_1 - y_2}{x_1 - x_2}(x - x_1)$$

$$y - 900 = \frac{1800 - 600}{3950 - 1350}(x - 2000)$$

$$y - 900 = \frac{1200}{2600}(x - 2000)$$

$$y - 900 = \frac{6}{13}(x - 2000)$$

$$13y - 11\ 700 = 6x - 12\ 000$$

$$0 = 6x - 13y - 300$$

iv. $0 = 6x - 13y - 300$

$0 = 6x - 13(2200) - 300$

$0 = 6x - 28\ 600 - 300$

$0 = 6x - 28\ 900$

$28\ 900 = 6x$

$4816.67 \doteq x$

\$4816.67

v. $0 = 6x - 13y - 300$

$0 = 6x - 13(3500) - 300$

$0 = 6x - 45\ 500 - 300$

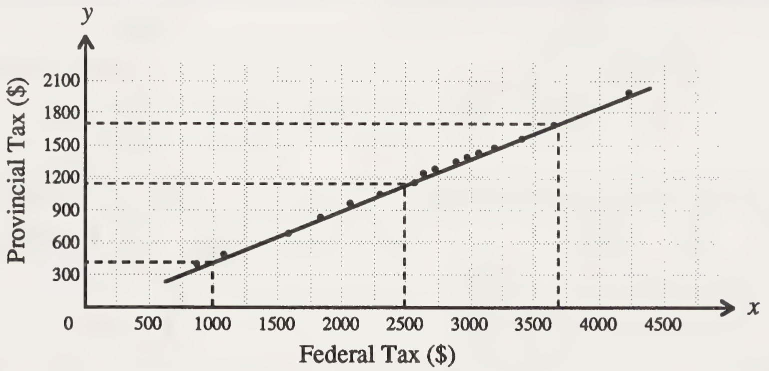
$0 = 6x - 45\ 800$

$45\ 800 = 6x$

$7633.33 \doteq x$

\$7633.33

Federal and Provincial Income Tax Payable in Northland Before Surtax, 1989



b. i. $\doteq$ \$450 ii. $\doteq$ \$1150 iii. $\doteq$ \$1700

iv.

$$0 = 6x - 13y - 300$$
$$0 = 6(300) - 13y - 300$$
$$0 = 1800 - 13y - 300$$
$$0 = 1500 - 13y$$
$$-1500 = -13y$$
$$115.38 \doteq y$$

\$115.38

v.

$$0 = 6x - 13y - 300$$
$$0 = 6(5500) - 13y - 300$$
$$0 = 33\ 000 - 13y - 300$$
$$0 = 32\ 700 - 13y$$
$$-32\ 700 = -13y$$
$$2515.38 \doteq y$$

\$2515.38

Activity 4

Determine the strength and type of correlation between the variables of a bivariate distribution.

1. a. They are the same.

   b. 0.34

   c. −0.75

   d. 0.86

2.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
18	3.18	15.5	3.14	2.5	0.04	0.10	6.25	0.0016
13	3.86	15.5	3.14	−2.5	0.72	−1.8	6.25	0.5184
12	2.44	15.5	3.14	−3.5	−0.70	2.45	12.25	0.49
15	2.72	15.5	3.14	−0.5	−0.42	0.21	0.25	0.1764
20	2.22	15.5	3.14	4.5	−0.92	−4.14	20.25	0.8464
18	3.19	15.5	3.14	2.5	0.05	0.125	6.25	0.0025
13	3.67	15.5	3.14	−2.5	0.53	−1.325	6.25	0.2809
9	2.92	15.5	3.14	−6.5	−0.22	1.43	42.25	0.0484
12	3.69	15.5	3.14	−3.5	0.55	−1.925	12.25	0.3025
10	4.29	15.5	3.14	−5.5	1.15	−6.325	30.25	1.3225
15	3.08	15.5	3.14	−0.5	−0.06	0.03	0.25	0.0036
13	3.25	15.5	3.14	−2.5	0.11	−0.275	6.25	0.0121
16	3.26	15.5	3.14	0.5	0.12	0.06	0.25	0.0144
16	3.43	15.5	3.14	0.5	0.29	0.145	0.25	0.0841
23	2.73	15.5	3.14	7.5	−0.41	−3.075	56.25	0.1681
18	3.41	15.5	3.14	2.5	0.27	0.675	6.25	0.0729
19	3.12	15.5	3.14	3.5	−0.02	−0.07	12.25	0.0004
13	3.32	15.5	3.14	−2.5	0.18	−0.45	6.25	0.0324
23	2.26	15.5	3.14	7.5	−0.88	−6.6	56.25	0.7744
17	2.91	15.5	3.14	1.5	−0.23	−0.345	2.25	0.0529
14	2.92	15.5	3.14	−1.5	−0.22	0.33	2.25	0.0484
14	3.14	15.5	3.14	−1.5	0	0	2.25	0
						−20.775	293.50	5.2533

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{-20.775}{\sqrt{293.5 \times 5.2533}}$$

$$r_{xy} \doteq -0.53$$

3.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
800	5.0	1200	8.2	−400	−3.2	1280	160 000	10.24
1100	7.5	1200	8.2	−100	−0.7	70	10 000	0.49
1300	7.8	1200	8.2	100	−0.4	−40	10 000	0.16
1800	12.3	1200	8.2	600	4.1	2460	360 000	16.81
900	6.2	1200	8.2	−300	−2.0	600	90 000	4
1400	10.4	1200	8.2	200	2.2	440	40 000	4.84
1100	8.2	1200	8.2	−100	0	0	10 000	0
1700	12.0	1200	8.2	500	3.8	1900	250 000	14.44
1200	9.6	1200	8.2	0	1.4	0	0	1.96
1000	8.5	1200	8.2	−200	0.3	−60	40 000	0.09
600	3.9	1200	8.2	−600	−4.3	2580	360 000	18.49
1200	8.2	1200	8.2	0	0	0	0	0
1500	9.8	1200	8.2	300	1.6	480	90 000	2.56
900	5.6	1200	8.2	−300	−2.6	780	90 000	6.76
						10 490	1 510 000	80.84

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{10\,490}{\sqrt{1\,510\,000 \times 80.84}}$$

$$r_{xy} \doteq 0.95$$

4.

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
820	381.20	2540	1183.30	– 1720	– 802.10	1 379 612	2 958 400	643 364.41
1060	493.60	2540	1183.30	– 1480	– 689.70	1 020 756	2 190 400	475 686.09
1540	716.40	2540	1183.30	– 1000	– 466.90	466 900	1 000 000	217 995.61
1780	828.80	2540	1183.30	– 760	– 354.50	269 420	577 600	125 670.25
2040	949.90	2540	1183.30	– 500	– 233.40	116 700	250 000	54 475.56
2280	1060.30	2540	1183.30	– 260	– 123.00	31 980	67 600	15 129.00
2560	1190.10	2540	1183.30	20	6.80	136	400	46.24
2620	1218.20	2540	1183.30	80	34.90	2 792	6 400	1 218.01
2740	1274.40	2540	1183.30	200	91.10	18 220	40 000	8 299.21
2860	1330.60	2540	1183.30	320	147.30	47 136	102 400	21 697.29
2980	1386.80	2540	1183.30	440	203.50	89 540	193 600	41 412.25
3020	1404.20	2540	1183.30	480	220.90	106 032	230 400	48 796.81
3160	1469.10	2540	1183.30	620	285.80	177 196	384 400	81 681.64
3340	1553.40	2540	1183.30	800	370.10	296 080	640 000	136 974.01
3660	1702.60	2540	1183.30	1120	519.30	581 616	1 254 400	269 672.49
4240	1972.90	2540	1183.30	1700	789.60	1 342 320	2 890 000	623 468.16
						5 946 436	12 786 000	2 765 587.03

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{5\ 946\ 436}{\sqrt{12\ 786\ 000 \times 2\ 765\ 587.03}}$$

$$r_{xy} \doteq 1.00$$

5. The correlation coefficient for question 4 is highest, so the Federal versus Provincial Income Tax for Northland is the set of bivariate data that has the highest correlation.

Activity 5

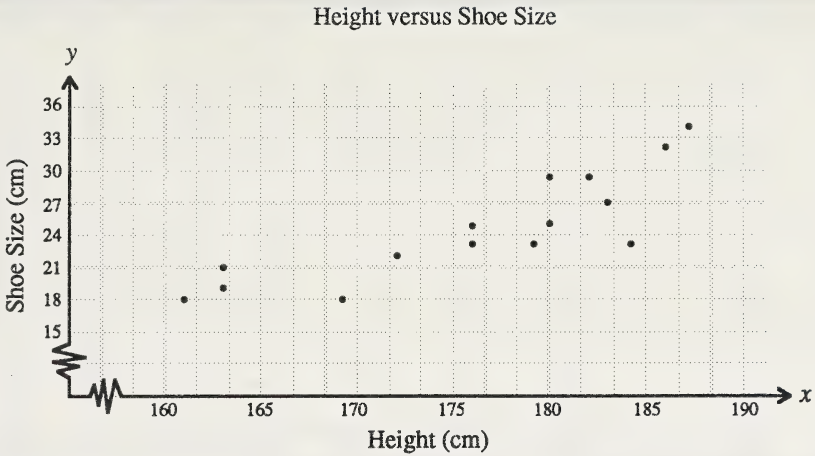
Collect, organize, and analyze sets of bivariate data.

The following solution is only an example of the components that you should have in your solution. Your solution should not be the same as this example.

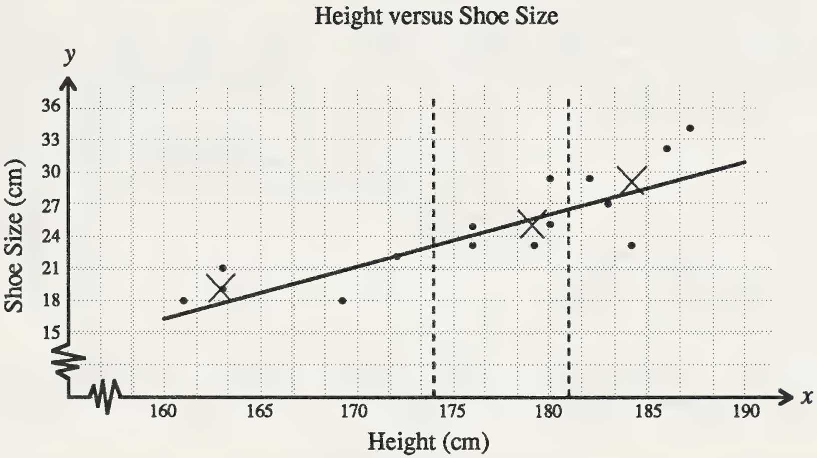
1. The chart of collected data is as follows:

Person	Height	Shoe Size
Eugene A.	180 cm	25 cm
Angie B.	163 cm	19 cm
Bobby C.	182 cm	29 cm
Debbie D.	176 cm	23 cm
Charlie E.	186 cm	32 cm
Tanya F.	163 cm	21 cm
Drago G.	176 cm	25 cm
Michele H.	161 cm	18 cm
Doug I.	180 cm	29 cm
Mai J.	172 cm	22 cm
Joe K.	184 cm	23 cm
Natalia L.	179 cm	23 cm
Kwok M.	183 cm	27 cm
Jolene N.	169 cm	18 cm
Guy O.	187 cm	34 cm

The scatterplot of the data is as follows:

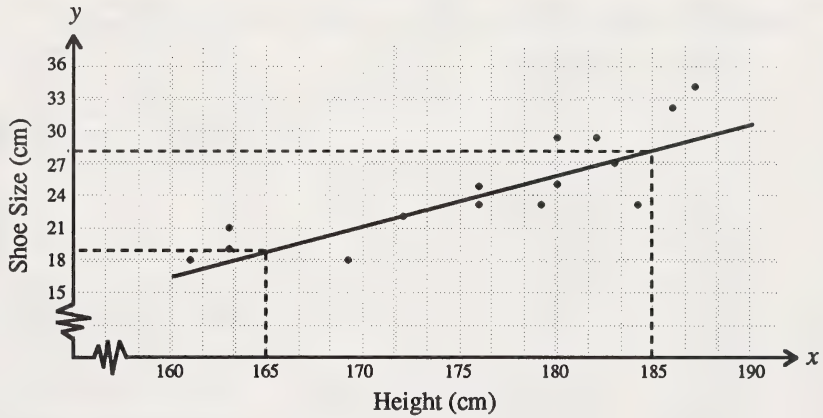


The line of best fit is as follows:



The equation of the line of best fit is as follows:

Height versus Shoe Size



From the graph, when the height is 185 cm, the shoe size is 28 cm. When the height is 165 cm, the shoe size is 19 cm. Use the points (185, 28) and (165, 19) to obtain the equation of the line.

$$y-y_2=\frac{y_1-y_2}{x_1-x_2}(x-x_2)$$

$$y-19=\frac{28-19}{185-165}(x-165)$$

$$y-19=\frac{9}{20}(x-165)$$

$$20y-380=9x-1485$$

$$0=9x-20y-1105$$

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
180	25	176	25	4	0	0	16	0
163	19	176	25	– 13	– 6	78	169	36
182	29	176	25	6	4	24	36	16
176	23	176	25	0	– 2	0	0	4
186	32	176	25	10	7	70	100	49
163	21	176	25	– 13	– 4	52	169	16
176	25	176	25	0	0	0	0	0
161	18	176	25	– 15	– 7	105	225	49
180	29	176	25	4	4	16	16	16
172	22	176	25	– 4	– 3	12	16	9
184	23	176	25	8	– 2	– 16	64	4
179	23	176	25	3	– 2	– 6	9	4
183	27	176	25	7	2	14	49	4
169	18	176	25	– 7	– 7	49	49	49
187	34	176	25	11	9	99	121	81
						497	1039	337

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

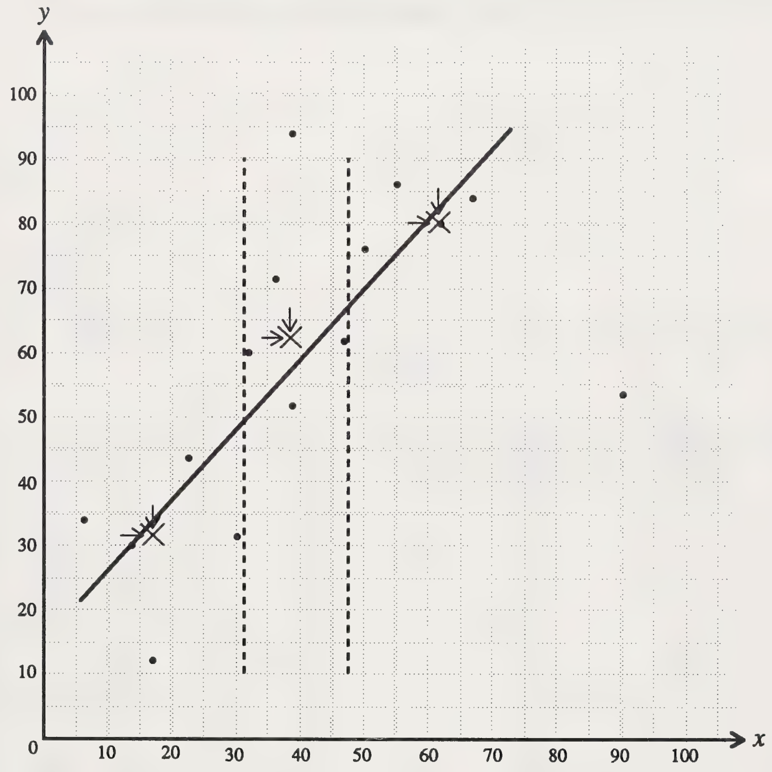
$$r_{xy} = \frac{497}{\sqrt{1039 \times 337}}$$

$$r_{xy} \doteq 0.84$$

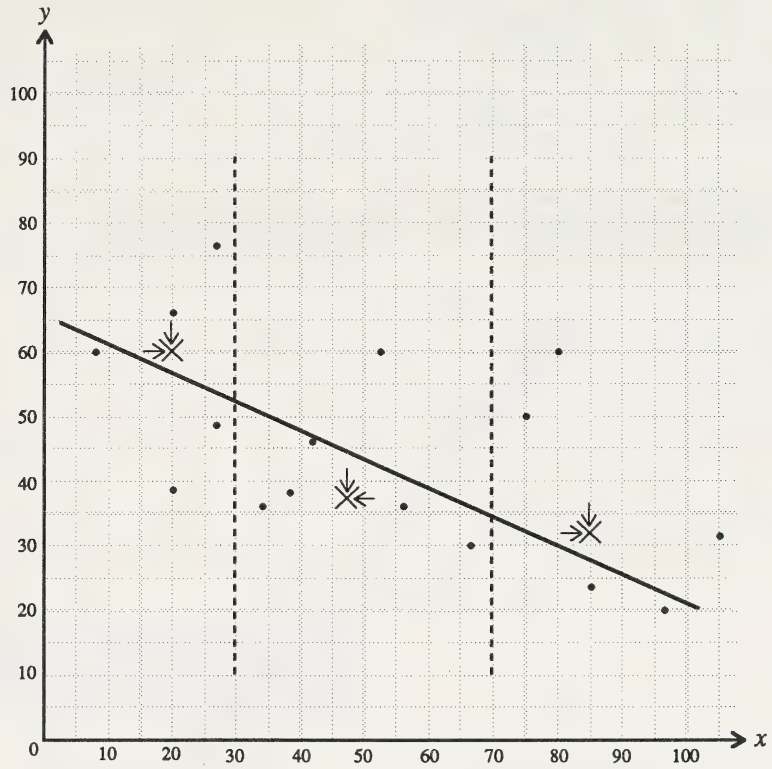
There is a definite positive correlation between the height of a person and the shoe size of that person. Neither of the factors is considered to be independent. Both factors are considered to be dependent on a third factor.

Extra Help

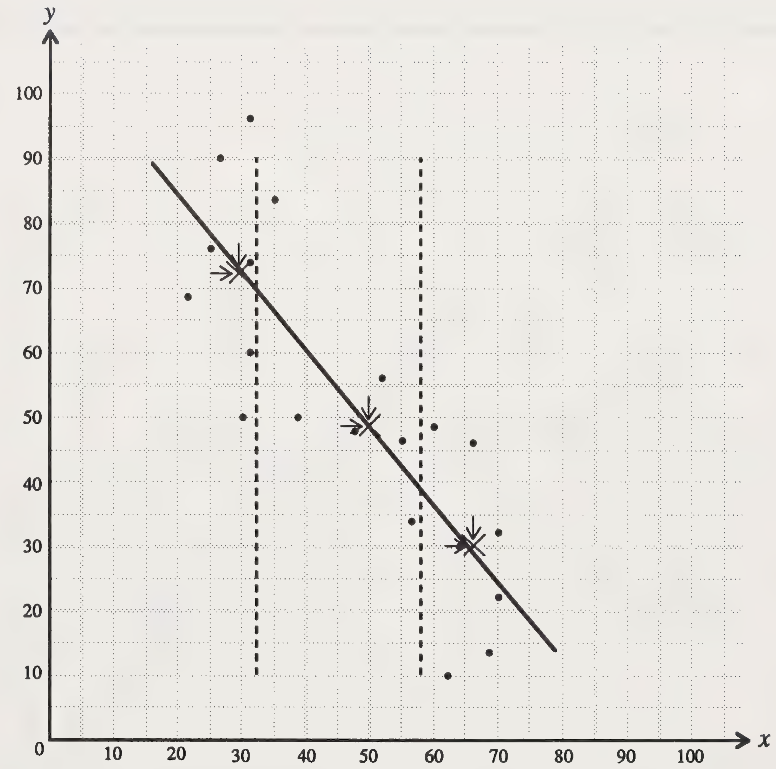
1.



2.



3.



Extensions

1. The three sums are as follows:

   – 2716.889
   – 52 639.11
   8017.6111

The correlation coefficient is $-0.132\,250$ which is rounded to -0.13.

This group of data has a weak negative correlation.

2. The three sums are as follows:

   1752.5
   4068.2778
   1528.5

   The correlation coefficient is 0.702 781 0 which rounds to 0.70.

   This is a positive correlation.

Exploring Topic 3

Activity 1

Design and administer a yes/no simple survey and collect and organize the results of the survey.

1. Answers will vary. The following are only some of many possible solutions.

   a. Do you prefer to watch hockey over baseball?

   b. Do you have a large breakfast in the morning?

   c. Do you prefer to drive large cars?

   d. Would you rather vacation in Europe than on a Pacific island?

2. Answers will vary. The following are only some of many possible solutions.

   a. Do you prefer to eat bananas over oranges?
   Do you prefer to eat oranges over apples?
   Do you prefer to eat apples over bananas?

   b. Would you prefer Brian Mulroney over Audrey McLaughlin to be the next prime minister?
   Would you prefer Audrey McLaughlin over Jean Chrétien to be the next prime minister?
   Would you prefer Brian Mulroney over Jean Chrétien to be the next prime minister?

3. Your answers will be near the following. The following are only some of many possible solutions.

   a. 4 heads

   b. The probability of getting 2 heads is $\frac{7}{64}$ or 0.109 375.

   The probability of getting 3 heads is $\frac{7}{32}$ or 0.218 75.

   The probability of getting 7 heads is $\frac{1}{32}$ or 0.031 25.

   The probability of getting 8 heads is $\frac{1}{256}$ or 0.003 906 25.

c. $\frac{93}{256}$ or 0.363 281

4. Answers will vary. The following is only one of many possible solutions.

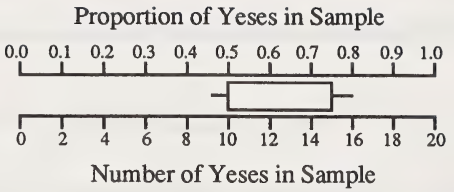
   The population percentage is 70% and the sample size is 20. Use a random number table selecting two digits at a time. If the selected number is 01 to 70, the response is yes. If the selected number is 00 or 71 to 99, the response is no. Fifty sets of twenty responses will be gathered.

Activity 2

Draw box and whisker plots of the results of multiple samples.

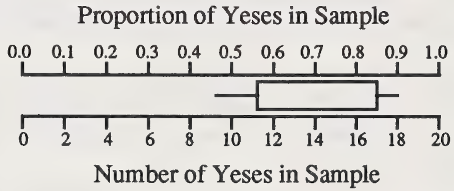
1. There can be a maximum of three values in the left whisker since 5% of 60 equals 3. The nine yeses have a frequency of 1; thus, start the left end of the left whisker at 9. The left side of the box has to start at ten yeses because the frequency for the ten yeses is 5, and these five values cannot be split. These five values cannot be placed in the left whisker, since the whisker can hold a maximum of three values and it already has one value. The frequency for fifteen yeses is 6 and the frequency for sixteen yeses is 2; thus, draw the right side of the box at fifteen yeses. Draw the right side of the right whisker at 16.

   This box and whisker plot has one value in the left whisker, fifty-seven values in the box, and two values in the right whisker.



2. The left whisker can have a maximum of five values since 5% of 100 equals 5. The nine yeses have a frequency of 1 and the ten yeses have a frequency of 4. Thus, these five values can be shown in the left whisker. Start the end of the left whisker at nine yeses and start the left side of the box at eleven yeses. The seventeen yeses have a frequency of 4 and the eighteen yeses have a frequency of 3. The three values for the eighteen yeses can be placed in the right whisker since it can hold a maximum of five values. The right side of the box ends at 17 and the right end of the right whisker ends at 18.

   This box and whisker plot has five values in the left whisker, ninety-two values in the box, and three values in the right whisker.



Activity 3

Use a 90% box and whisker plot chart to find the confidence interval for a survey result.

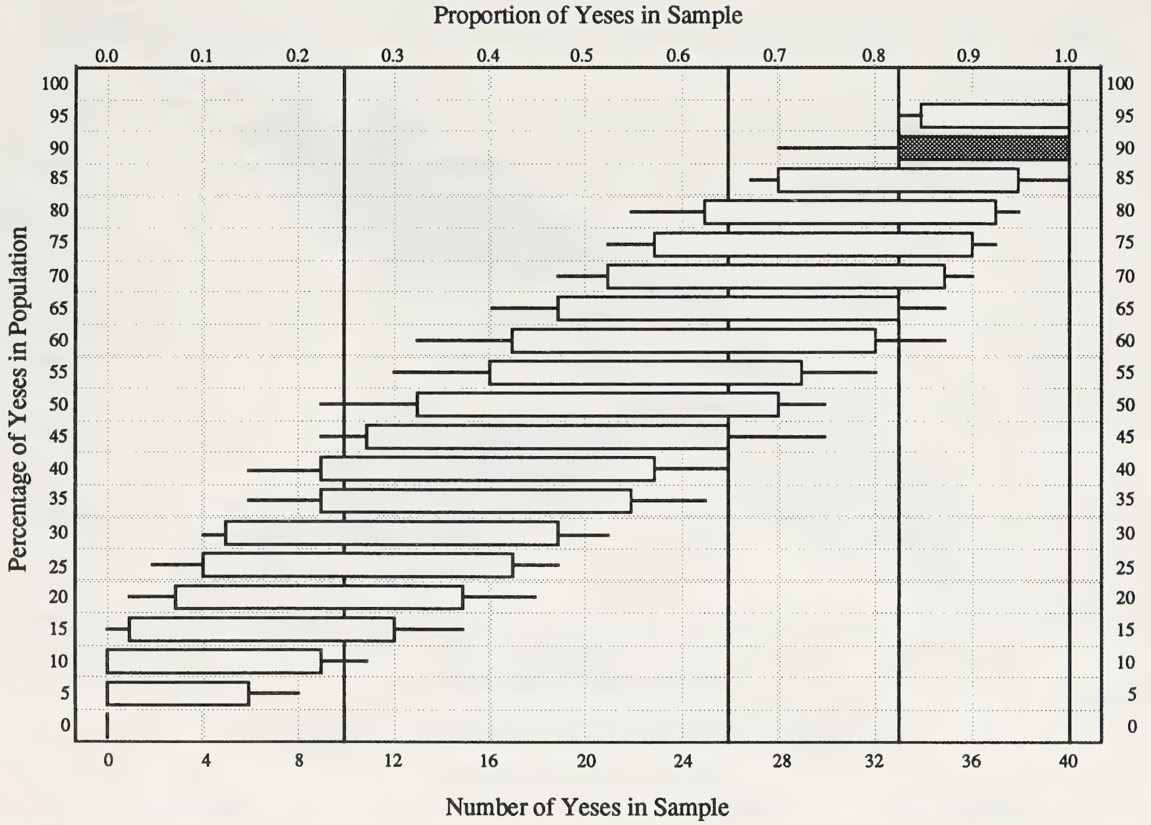
1. a. A sample proportion of 0.25 is unlikely.

   b. A sample proportion of 0.65 is unlikely.

   c. A sample proportion of 0.825 is likely.

   d. A sample proportion of 1.00 is likely.

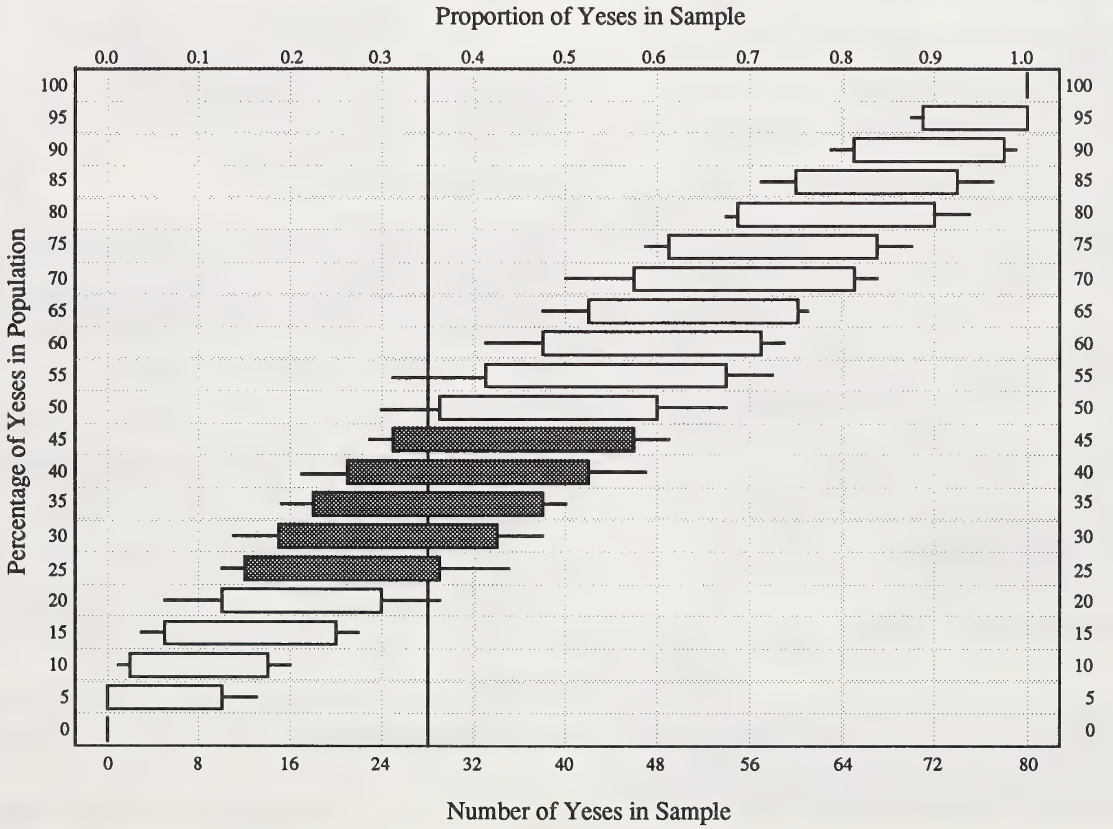
90% Box and Whisker Plots from Samples of Size 40



2. a. unlikely b. unlikely c. likely d. likely e. likely

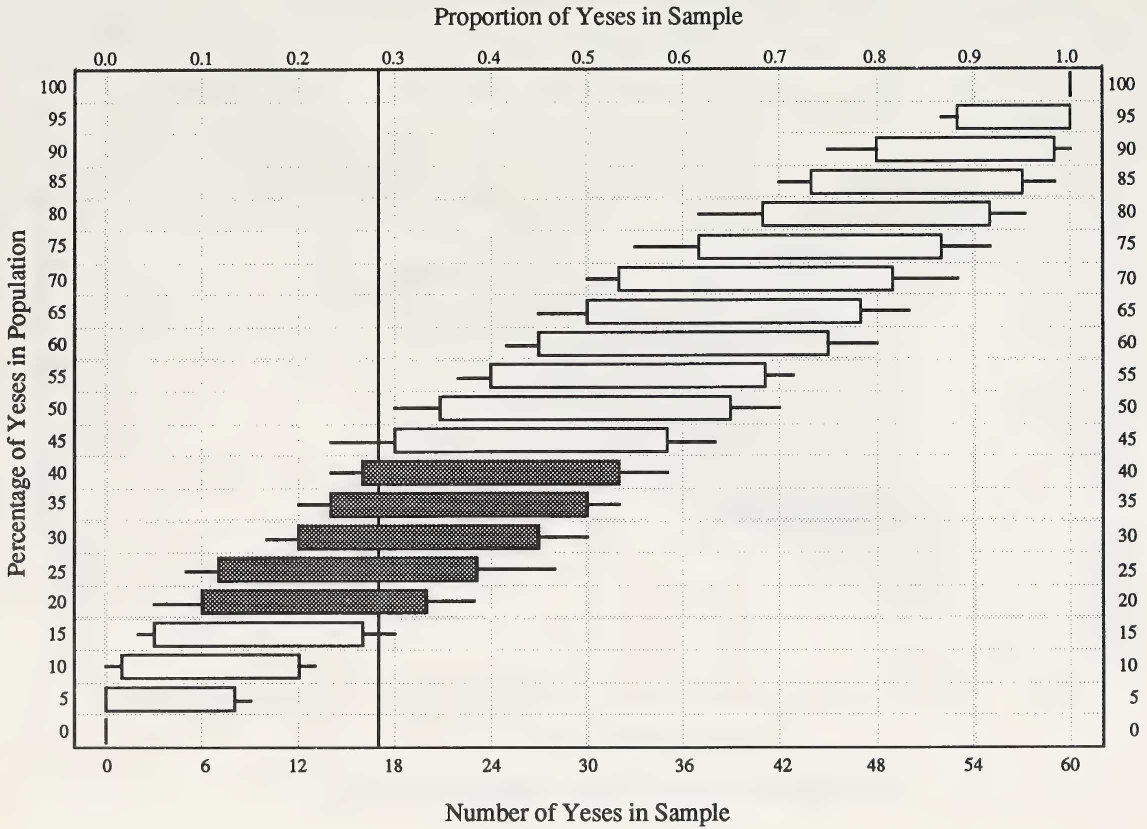
f. likely g. likely h. unlikely i. unlikely j. unlikely

90% Box and Whisker Plots from Samples of Size 80



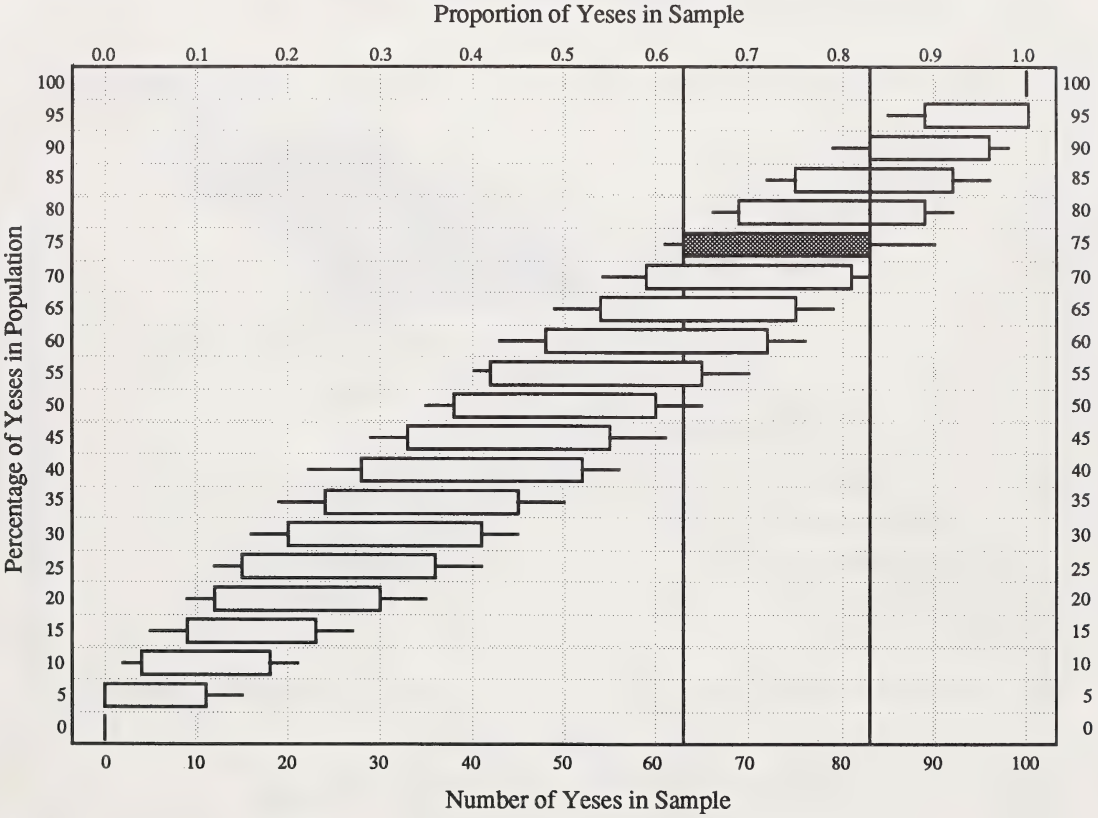
3. This is likely for the population percentages of 20% to 40% inclusive.

90% Box and Whisker Plots from Samples of Size 60



4. The sample proportions will go from 0.63 to 0.83 inclusive.

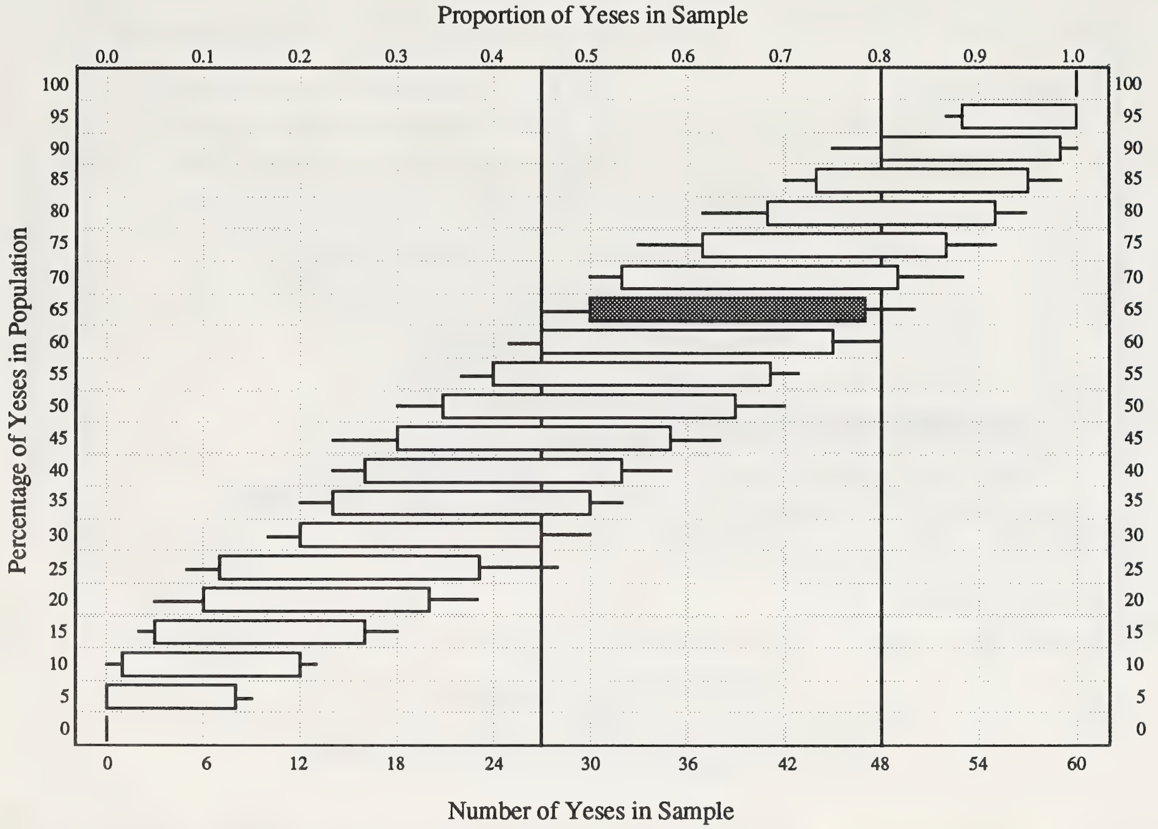
90% Box and Whisker Plots from Samples of Size 100



5. a. A sample proportion of 0.45 is unlikely.

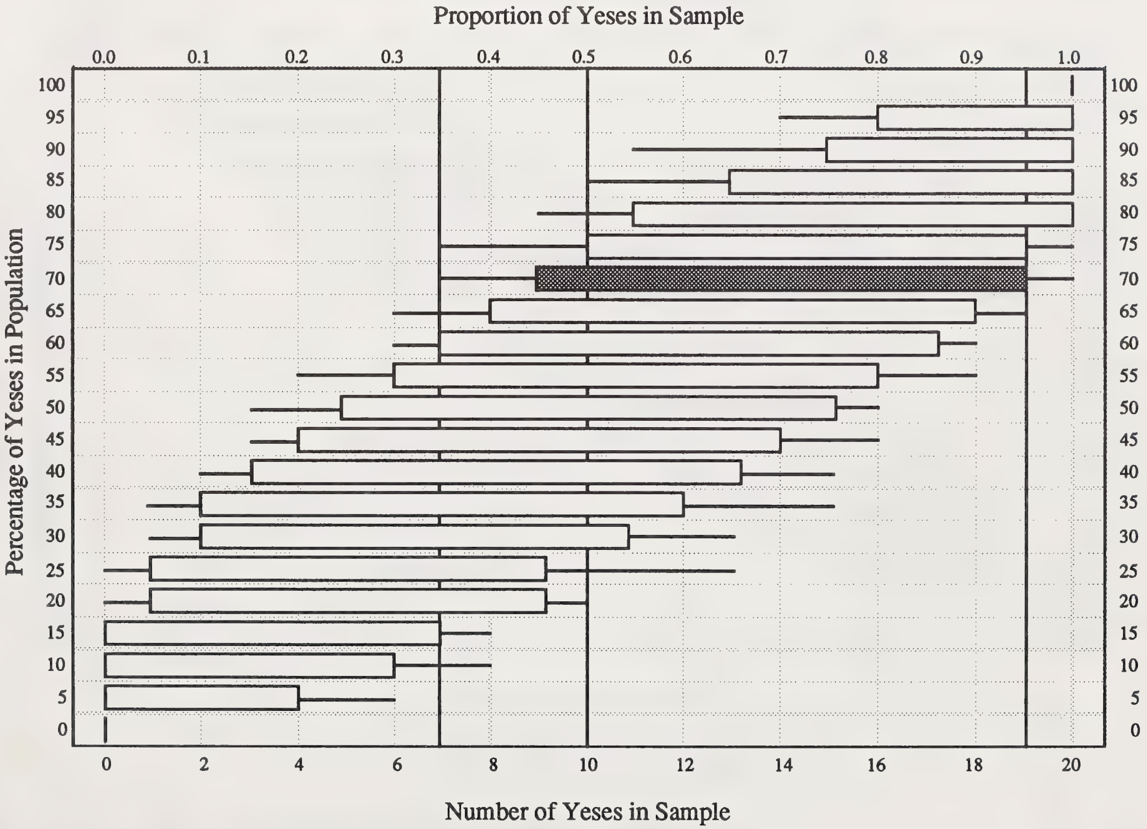
b. A sample with forty-eight people having alcohol in their blood is unlikely.

90% Box and Whisker Plots from Samples of Size 60



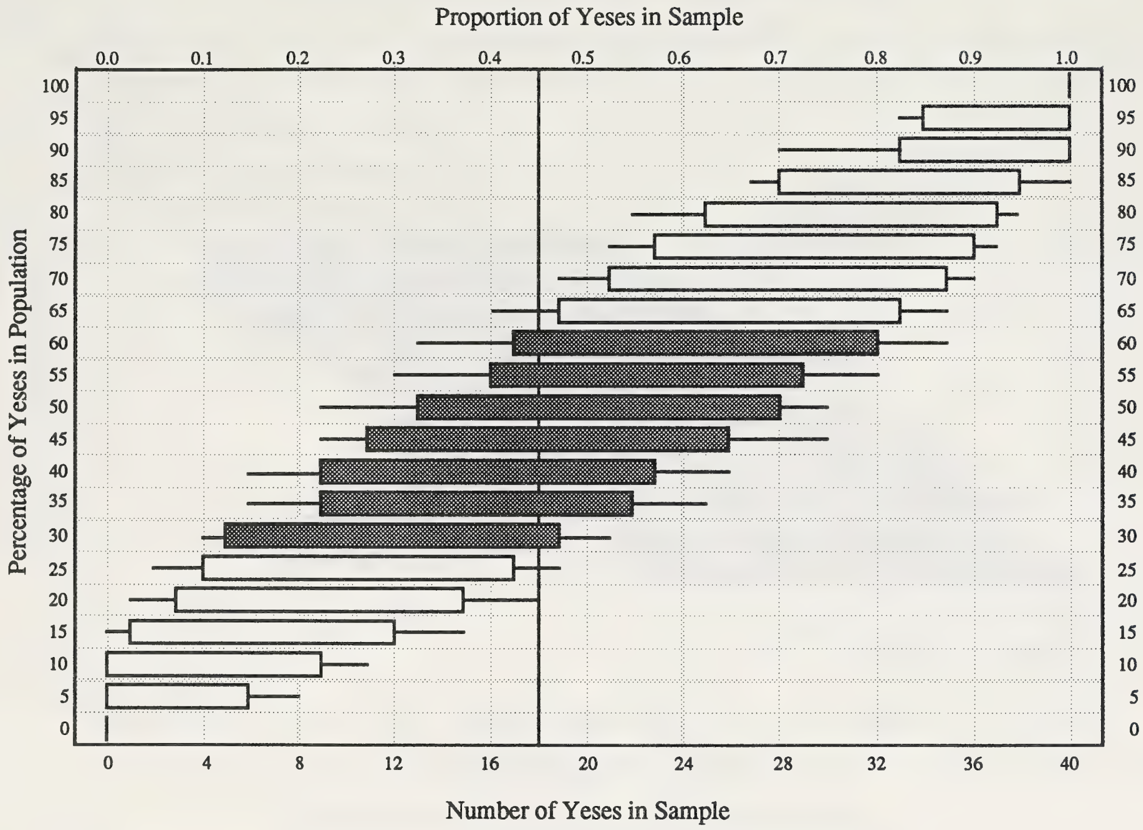
6. a. It is likely that a sample proportion of 0.95 of the people will approve of the tax.
   b. It is likely that ten people will approve of the tax.
   c. It is unlikely that a sample proportion of 0.35 of the people will approve of the tax.

90% Box and Whisker Plots from Samples of Size 20



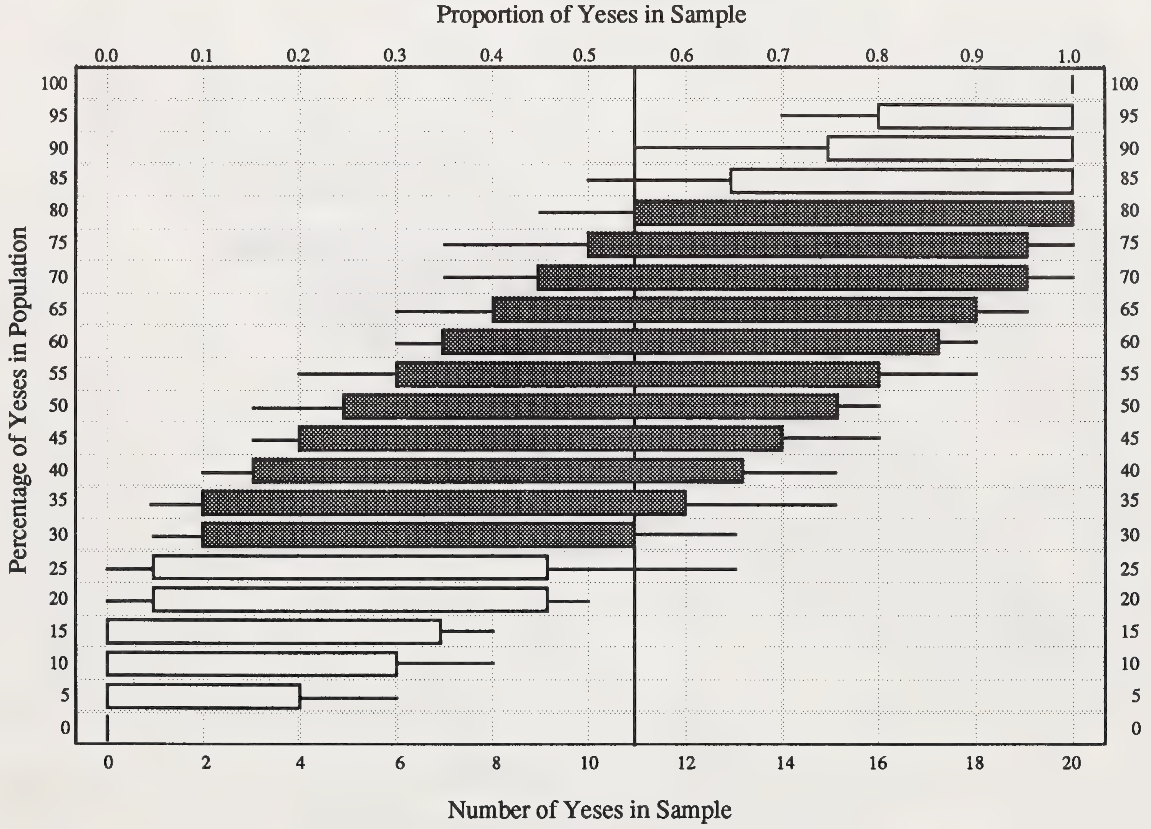
7. Thirty percent to sixty percent of the people will buy the car.

90% Box and Whisker Plots from Samples of Size 40



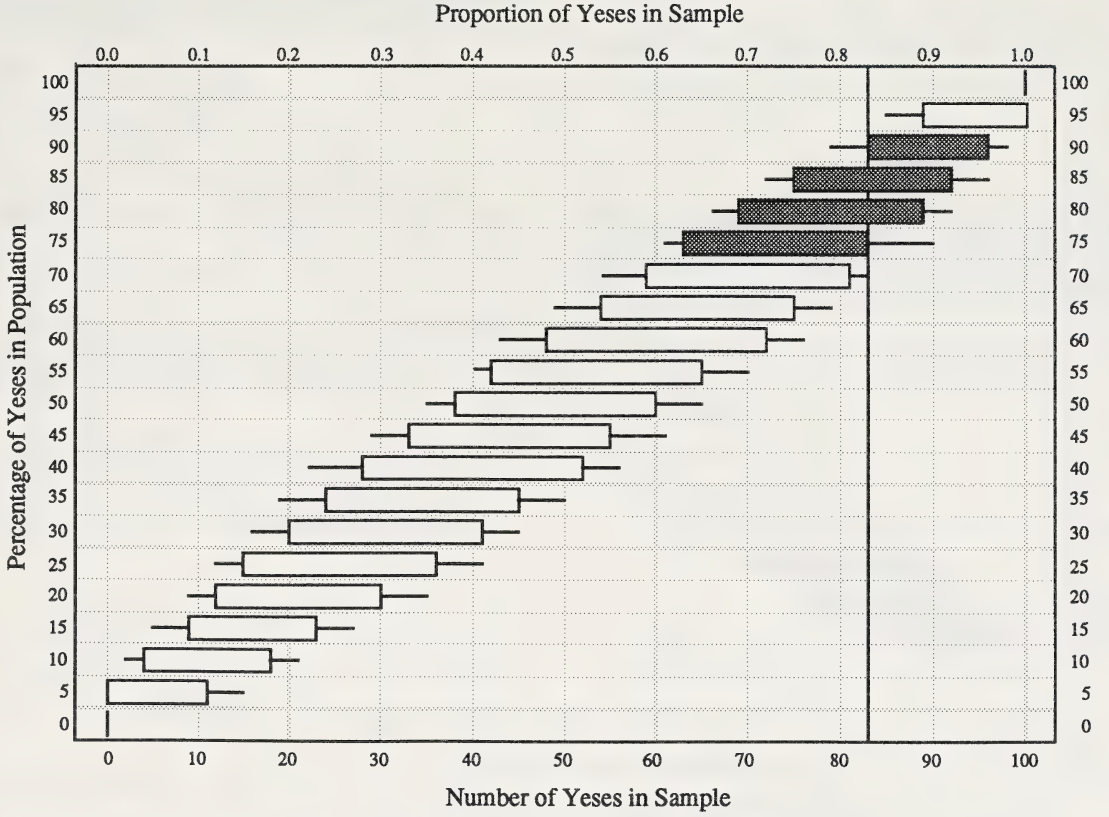
8. a. The confidence interval is from 30% to 80% inclusive.

90% Box and Whisker Plots from Samples of Size 20



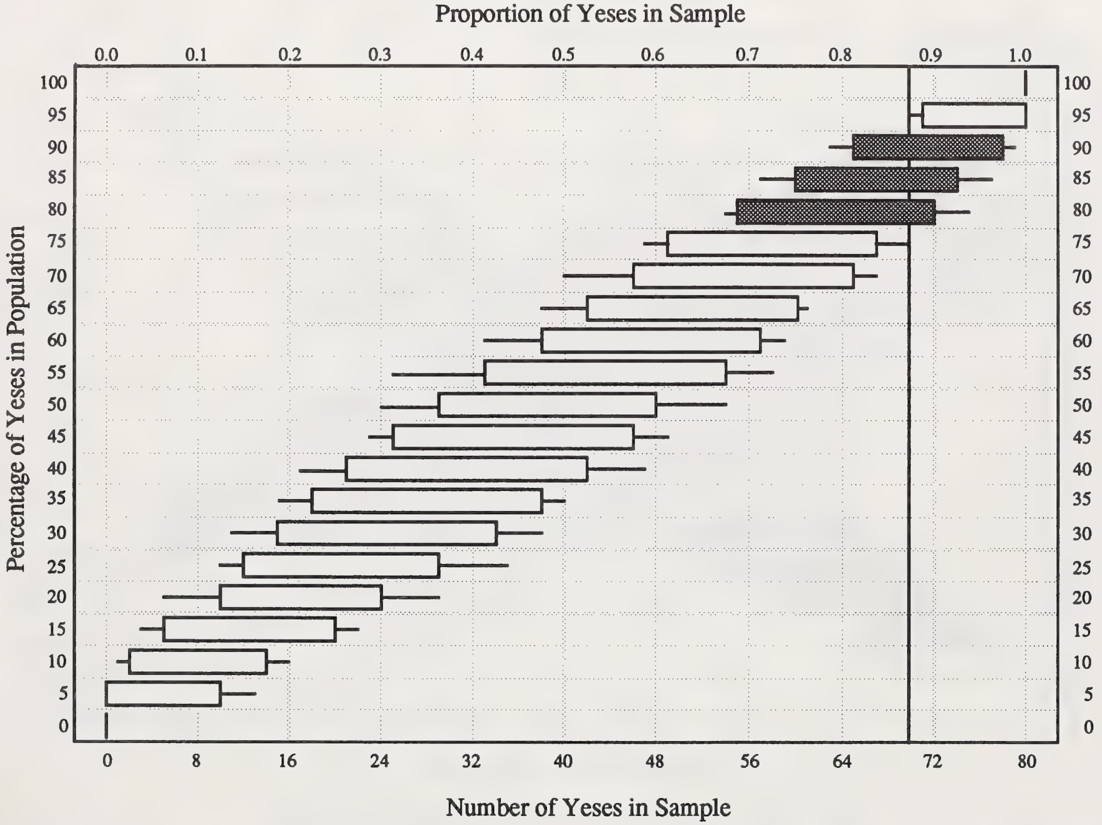
b. The confidence interval is from 75% to 90% inclusive.

90% Box and Whisker Plots from Samples of Size 100



c. The confidence interval is from 80% to 90% inclusive.

90% Box and Whisker Plots from Samples of Size 80



Activity 4

Draw statistical conclusions and make inferences to populations, and explain the confidence with which such conclusions and inferences are made based on the results of yes/no surveys.

1. 0.70 is not a likely sample proportion.

2. 90% of the samples will be likely.
   90% of 500
   $0.90 \times 500 = 450$
   450 of the population percentages will be likely.

3. The sample size should be decreased.

4. From the box and whisker plots for a sample size of 60, the 90% confidence interval will be 70% to 90%. A vertical line from 49 intersects the 70%, 75%, 80%, 85%, and 90% boxes.

   70% of 60 is 42 and 90% of 60 is 54. If the sixty people of a random sample of this population were asked if they ate breakfast every morning, then 42 to 54 people would respond yes. These results will be true for 90 samples out of every 100 samples.

5. From the box and whisker plots for a sample size of 40, the 90% confidence interval will be 25% to 55%. A vertical line from 17 intersects the 25%, 30%, 35%, 40%, 45%, 50%, 55% and 60% boxes. 25% of 40 is 10 and 60% of 40 is 24.

   If the forty people of a random sample of this population were asked if they favoured running the public school year-round, then 10 to 22 people would respond yes. These results will be true for 90 samples out of every 100 samples.

6. From the box and whisker plots for a sample size of 80, the 90% confidence interval will be 85% to 95%. A vertical line from 73 intersects the 85%, 90%, and 95% boxes.
   85% of 80 is 68 and 95% of 80 is 76.

   If the eighty people of a random sample of this population were asked if they wanted more parks, 68 to 76 people would respond yes. The results will be true for 90 samples out of every 100 samples.

Activity 5

Assess the strengths, weaknesses, and biases of given samples.

1. a. Yes. Assuming that every student has a phone and that the phone numbers are assigned randomly, this is a random sampling of students.

   b. Yes.

   c. No. Students will not be selected independently. Entire families may be selected.

2. This situation may or may not be a random selection. It is impossible to tell from the results of the selection whether a random selection has been performed.

3. It would be impossible to get a complete and up-to-date list of all of the people that live in Calgary. Since you will not be able to list all of the people that live in Calgary, all of the people will not have the same chance of being selected.

4. a. self-selected

   b. stratified random

   c. judgement

5. Answers will vary. Some possible solutions are as follows:

   - People may be giving answers that they believe are socially acceptable.

   - People may be trying to give the answers that they believe the interviewer wants to hear.

   - People may be trying to seem knowledgeable by trying to answer questions that they do not understand.

   - People often do not correctly remember numbers.

6. Answers will vary. Some possible solutions are as follows:

   - People in the armed forces or in institutions may not have been surveyed.

   - Women may consider it more acceptable to be considered married and thus, report themselves as being married.

7. Answers will vary.

Extra Help

1. Note that the diagram is at the end of question 5.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	55	0.055	Box
1	0.05	149	0.149	
2	0.10	166	0.166	
3	0.15	215	0.215	
4	0.20	169	0.169	
5	0.25	132	0.132	
6	0.30	59	0.059	
7	0.35	47	0.047	
8	0.40	8	0.008	Right Whisker
9	0.45	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
20	1.00	0	0	
Total		1000	1.000	

2.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	17	0.017	Left Whisker
1	0.05	59	0.059	Box
2	0.10	136	0.136	
3	0.15	183	0.183	
4	0.20	170	0.170	
5	0.25	141	0.141	
6	0.30	103	0.103	
7	0.35	77	0.077	
8	0.40	63	0.063	
9	0.45	49	0.049	
10	0.50	2	0.002	Right Whisker
11	0.55	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
20	1.00	0	0	
Total		1000	1.000	

3.

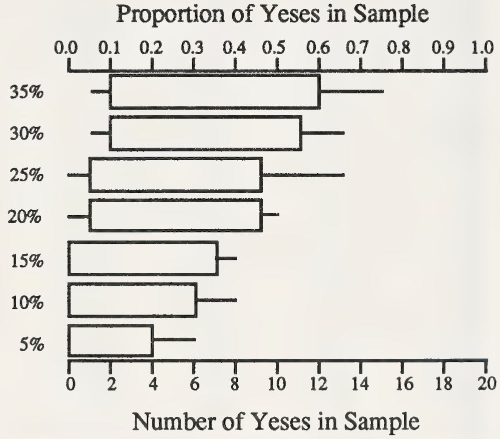
Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	4	0.004	Left Whisker
1	0.05	47	0.047	Box
2	0.10	82	0.082	
3	0.15	117	0.117	
4	0.20	151	0.151	
5	0.25	157	0.157	
6	0.30	138	0.138	
7	0.35	117	0.117	
8	0.40	101	0.101	
9	0.45	49	0.049	
10	0.50	20	0.020	Right Whisker
11	0.55	11	0.011	
12	0.60	5	0.005	
13	0.65	1	0.001	
14	0.70	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
20	1.00	0	0	
Total		1000	1.000	

4.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	0	0	
1	0.05	22	0.022	Left Whisker
2	0.10	39	0.039	Box
3	0.15	54	0.054	
4	0.20	93	0.093	
5	0.25	107	0.107	
6	0.30	123	0.123	
7	0.35	140	0.140	
8	0.40	134	0.134	
9	0.45	108	0.108	
10	0.50	83	0.083	
11	0.55	57	0.057	
12	0.60	32	0.032	Right Whisker
13	0.65	8	0.008	
14	0.70	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
20	1.00	0	0	
Total		1000	1.000	

5.

Number of Yeses	Sample Proportion	Frequency	Proportion of All Trials	
0	0.00	0	0	
1	0.05	7	0.007	Left Whisker
2	0.10	52	0.052	Box
3	0.15	78	0.078	
4	0.20	102	0.102	
5	0.25	111	0.111	
6	0.30	123	0.123	
7	0.35	129	0.129	
8	0.40	113	0.113	
9	0.45	98	0.098	
10	0.50	76	0.076	
11	0.55	57	0.057	
12	0.60	42	0.042	
13	0.65	11	0.011	Right Whisker
14	0.70	0	0	
15	0.75	1	0.001	
16	0.80	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
20	1.00	0	0	
Total		1000	1.000	



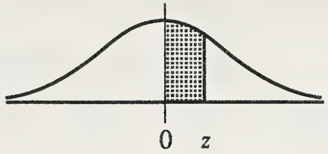
Extensions

1. Answers will vary.

2.
```
100 CLEAR
110 DIM N(60), S(60, 20), S$(60, 20)
120 FOR A = 1 TO 60
130 FOR B = 1 TO 20
140 S(A, B) = INT(100*RND (1))
150 IF S(A, B) < 43 THEN S$(A, B) = "Y"
160 IF S(A, B) >= 43 THEN S$(A, B) = "N"
170 IF S(A, B) < 43 THEN N(A) = N(A) + 1
180 NEXT B
190 NEXT A
195 PR#1
200 FOR A = 1 TO 60
210 PRINT A;"  ";
220 FOR B = 1 TO 20
230 PRINT S$(A, B);
240 NEXT B
250 PRINT "  "; N(A)
260 NEXT A
265 PR#0
```

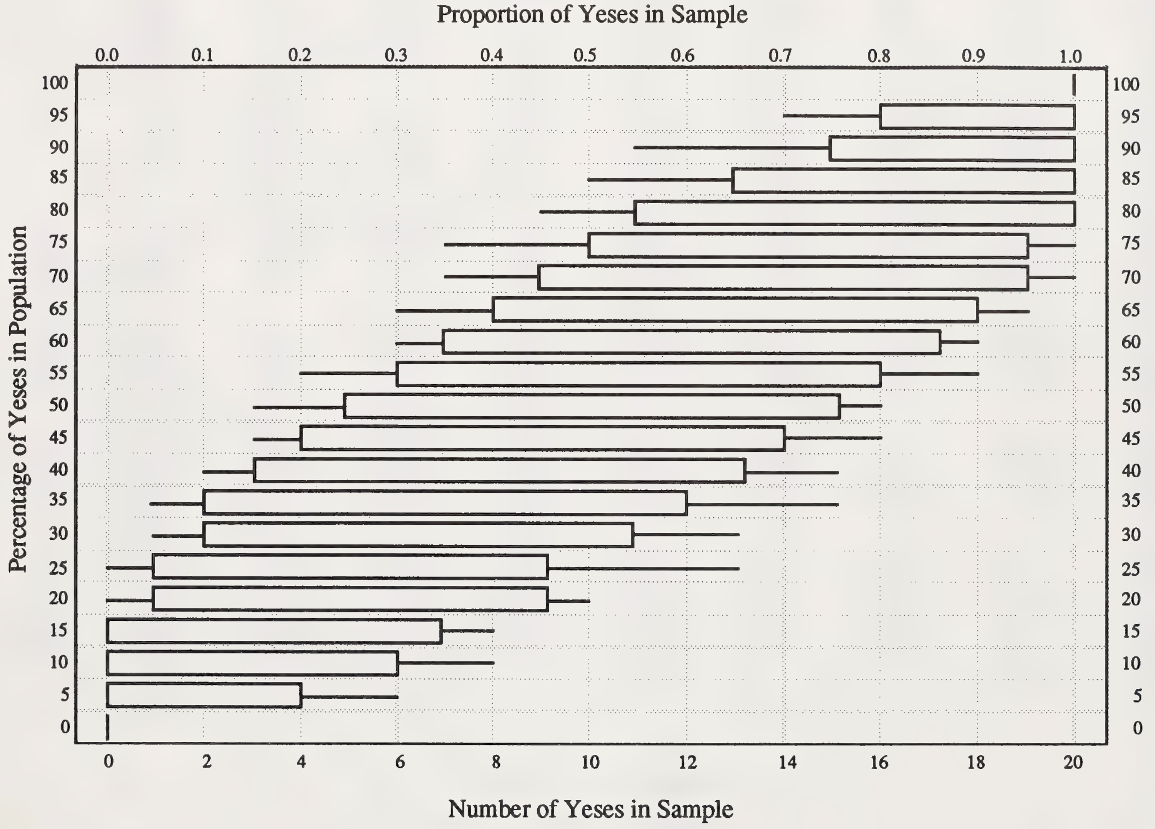
Appendix B
Charts and Tables

The Standard Normal Distribution Table

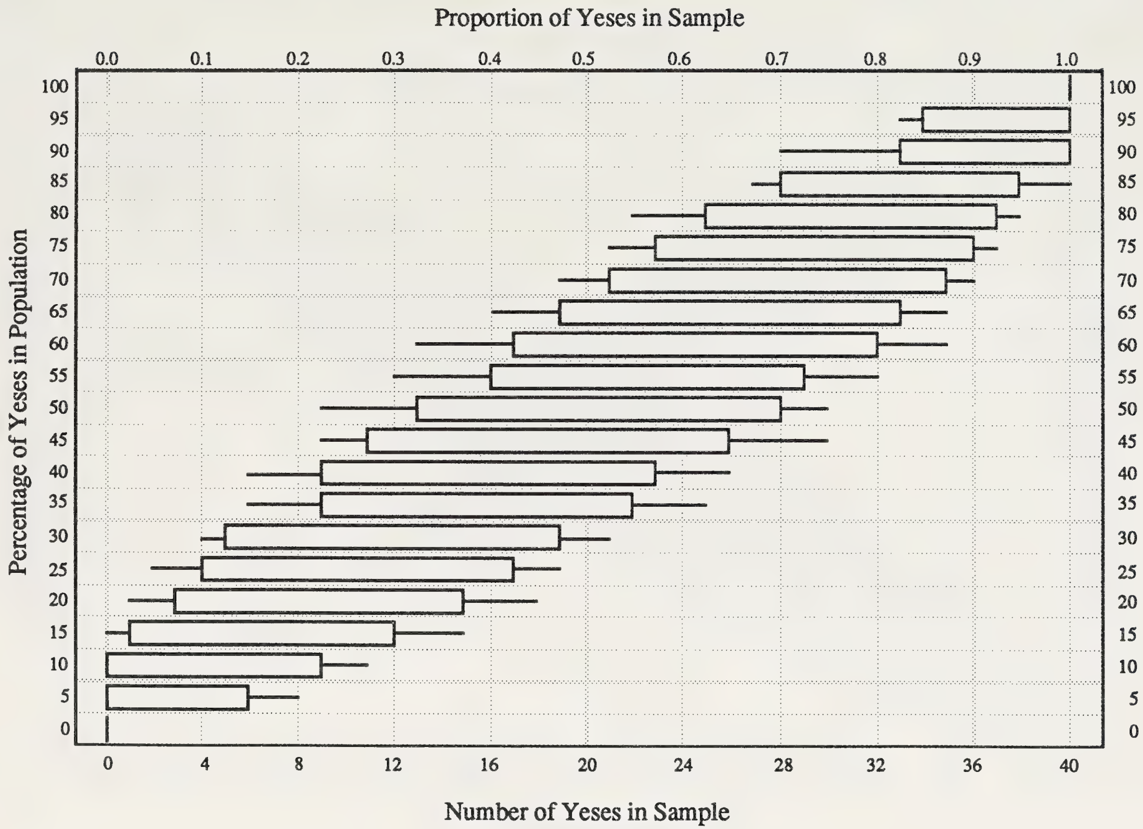


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

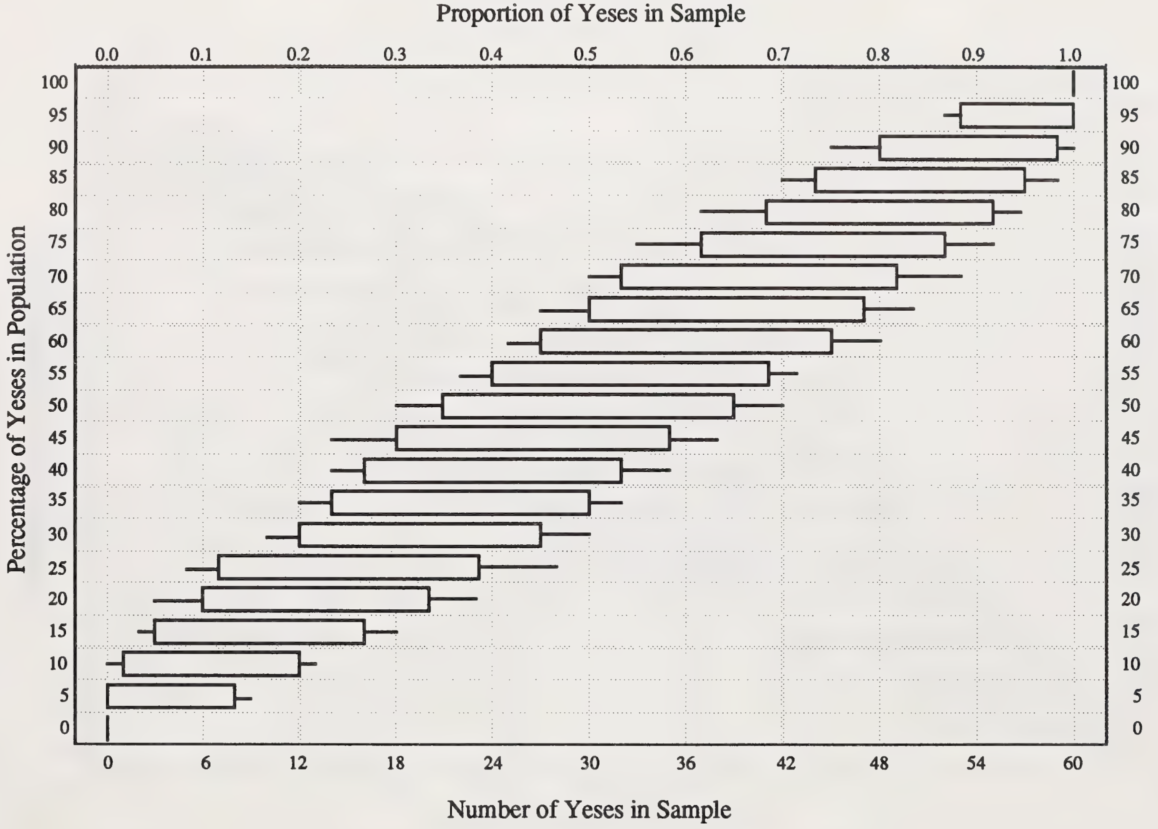
90% Box and Whisker Plots from Samples of Size 20



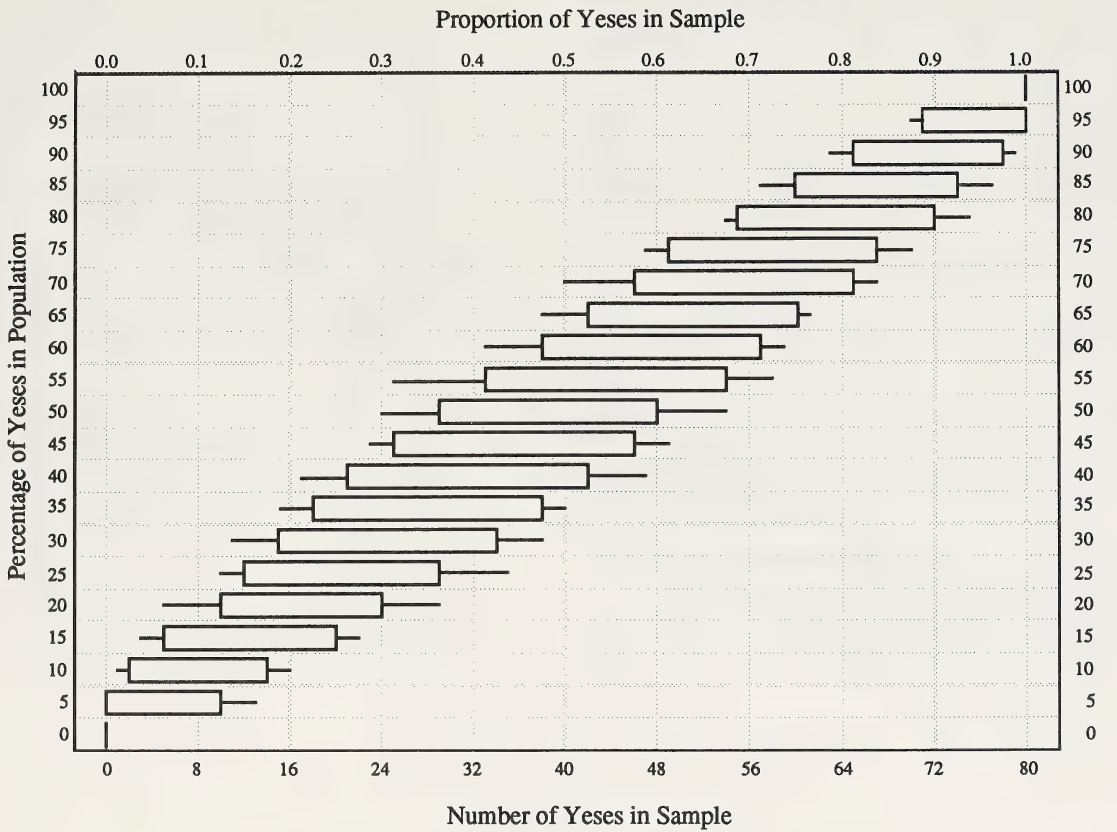
90% Box and Whisker Plots from Samples of Size 40

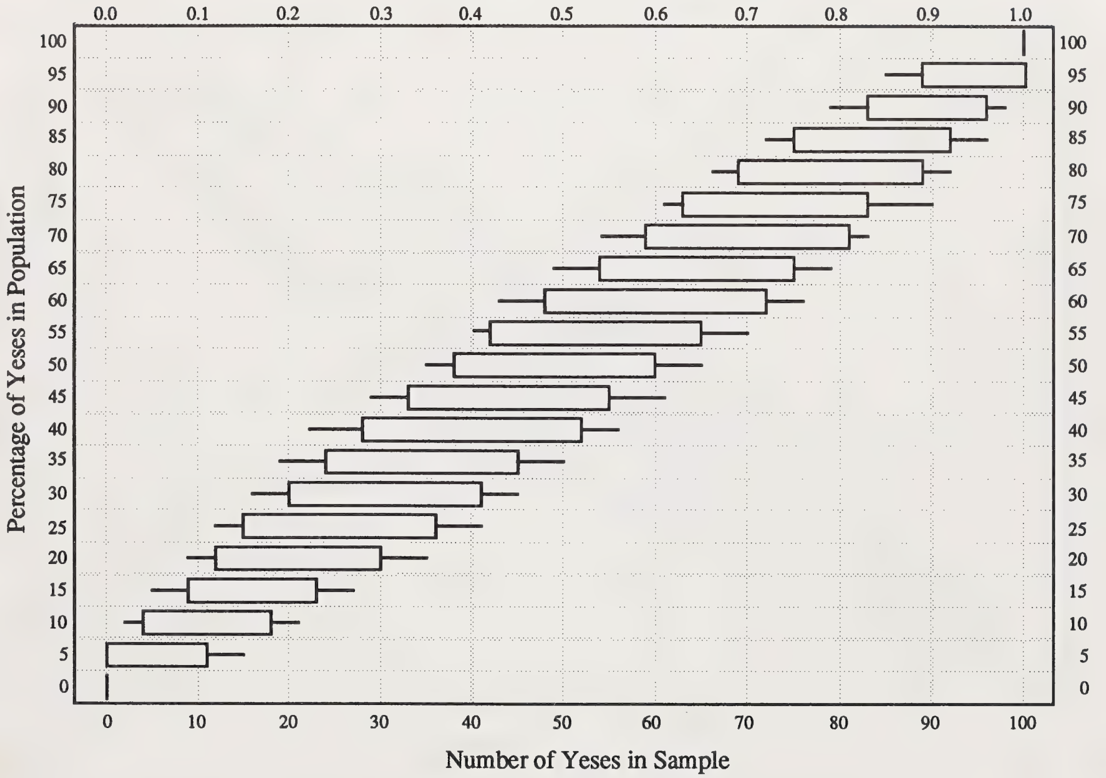


90% Box and Whisker Plots from Samples of Size 60



90% Box and Whisker Plots from Samples of Size 80





90% Box and Whisker Plots from Samples of Size 100

Random Number Table

```
91354 55254 27245 38952 64544 74382 78774 82633 63700 11314     68077 55470 47947 75295 72665 61065 98082 49134 03419 36994
89780 09210 30185 98209 95035 20617 39446 64982 36050 98573     90730 34887 80069 30309 65984 68523 25963 77223 96459 88451
27573 59612 00134 87398 42862 93457 84463 01818 22380 11567     57121 35920 88850 30834 96305 44370 10497 49903 12074 03141
85137 59784 93526 87997 11242 40282 00593 46117 23918 66144     40947 63264 82854 83210 48517 22220 85993 58779 91855 07072
56332 03926 00418 11864 18511 44365 23169 36409 91484 10628     11829 65082 46505 47443 02753 81203 56467 67581 17265 37885

93838 41029 52745 26477 72391 52073 04500 41550 41159 28801     10761 56677 19045 50345 44989 64176 91007 76083 67896 20088
50220 01681 93898 96194 96210 44203 49853 02769 78170 18268     74370 97699 12936 80303 32713 30701 14833 96903 41043 56321
92818 03390 12757 30157 35202 67082 74146 13592 07090 09843     11141 13473 99854 79327 68809 85414 74192 10832 35745 25258
95596 25049 70303 52454 96009 64069 57465 03297 31494 18575     57714 76111 17373 72975 61754 47411 78203 91292 29386 01766
36063 66160 71932 35178 97509 73058 02532 14495 71203 42038     68386 86001 74483 03805 75282 51273 45781 13568 00022 39970

50401 31459 63598 59996 36102 80631 22257 90489 50216 46515     15943 19318 59064 58960 95619 95354 10766 89425 72510 84821
24921 06543 54909 98869 93570 48474 10488 13014 33482 06693     06941 32177 75432 07525 69419 77968 63375 92745 91617 02807
26970 15155 97300 29520 40641 52683 74093 56001 15617 89499     39567 37044 65092 88266 90539 87084 31110 66798 50182 47436
03418 73715 70410 26314 14382 01588 29490 45973 89707 97708     57207 92047 88816 13394 38751 34160 27247 28788 25502 00101
99846 83760 54789 92135 72547 22713 91763 44714 49308 13937     75834 68796 73181 78964 98546 47753 20025 12283 02032 51970

78565 27886 94053 21915 52995 34506 09472 94823 01776 10648     88356 89347 49675 97541 11138 67509 00662 13160 83544 17504
38589 59840 74622 13424 91323 43460 23297 93790 96048 16809     74041 47929 39515 93757 37788 04174 53324 29696 22865 95255
65678 04367 76160 65951 74818 32244 51635 87207 93009 56798     66141 58067 62287 28381 37590 56245 55181 12419 22316 16544
65594 07462 70842 32503 35867 29481 31272 96743 31550 02222     55850 90711 71546 41536 87834 98883 59651 64354 82977 47660
58682 02625 74574 16142 12940 76790 66263 23521 68466 02228     35709 94015 45483 30674 00069 85725 40983 57283 05911 36194

38899 83762 78294 26089 32283 33713 63133 92494 95298 82975     10597 10379 85808 03948 09996 29201 03293 54644 73461 17149
25431 00352 12471 34215 88435 86168 55537 28478 60258 31856     29416 31377 02954 74594 44311 49999 82170 56695 63754 19427
60000 15504 22937 97555 26697 49032 46497 79438 48357 07122     15609 56207 90731 86958 92633 11295 91996 08578 76527 38347
28015 92526 61170 37459 55681 78499 21341 27672 97859 38511     99635 44279 18569 89785 37629 78369 97084 36882 73415 64786
06661 00642 17154 91001 43063 34820 09471 54505 56189 09531     81512 77567 94451 97406 21469 69244 09775 05362 60808 86602
```

Mathematics 30

9MA30P36

L.R.D.C.
Producer

THIRD EDITION
1991